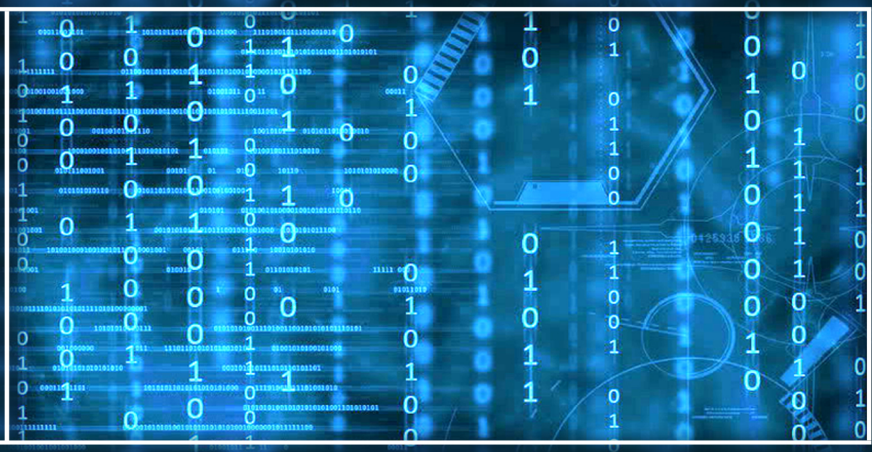
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 12 Issue 2
February 2021



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

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 12 Issue 2 February 2021
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Model-driven Framework for Requirement Traceability

Nader Kesserwan[1], Jameela Al-Jaroodi[2]
School of Engineering, Mathematics, and Science (SEMS)
Robert Morris University
Pittsburgh, USA

***Abstract*—In software development, requirements traceability is often mandated. It is important to apply to support various software development activities like result evaluation, regression testing and coverage analysis. Model-Driven Testing is one approach to provide a way to verify and validate requirements. However, it has many challenges in test generation in addition to the creation and maintenance of traceability information across test-related artifacts. This paper presents a model-based methodology for requirements traceability that relies on leveraging model transformation traceability techniques to achieve compliance with DO-178C standard as defined in the software verification process. This paper also demonstrates and evaluates the proposed methodology using avionics case studies focusing on the functional aspects of the requirements specified with the UCM (Use Case Maps) modeling language.**

***Keywords*—Requirements; traceability; model transformation; do-178c; model-driven testing; traceability scheme**

I. Introduction

The largest part of traceability research so far has been done in the last two decades by the requirements engineering community [1]. Traceability, known as the ability to describe and follow the life of software artifacts [2], has become more important and traceability topics are being researched in many other areas of software development. One example is model-driven development where some components of the software development process are executed automatically using model transformations [3]. Model-driven development provides the foundation for the use of models as primary artefacts throughout the software development phases [4]. The variety of different models produced in the model-driven process pose challenges to requirements traceability and assessment. This diversity of artifacts results in an intricate relationship between requirements and the various models. The model-based testing (MBT) is a technique where test cases are generated from models [5]. MBT needs the ability to relate the "abstract values of the specification to the concrete values of the implementation" [6]. The relationships between artifacts play an important role to support automation of testing activities and it has been recognized for some time [7]. Relationships between behavioral models and test cases and between test cases and test results support better capabilities to measure coverage, evaluate results and perform selective regression testing. As a result, creating and maintaining explicit relationships among test-related artifacts is a main challenge to the automated support of these activities.

In this paper, model transformation techniques are used to create traceability links among MBT artifacts during the test generation process. The approach extends previous testing methodology presented in [8] that generates tests based on behavioral models. This paper's contribution is building a traceability model to support the creation and persistence of such relationships among heterogeneous models representing various testing artifacts. Moreover, this work enables the support for traceability visualization, model-based coverage analysis and result evaluation. The case study used in this work is an industrial product, flight management system (FMS), to evaluate the correctness of the approach that ensures all the generated test cases determine correctly the behavior of the FMS and are traceable to requirements.

The rest of this paper is organized as follows. Section 2 offers background information on traceability and model-based approaches in requirements and testing. A discussion of some related work about model transformation, model-based test generation, and traceability applied to automated testing approaches is presented in Section 3. Section 4 presents and describes the proposed approach, which is followed by Section 5 where two case studies are used to demonstrate the applicability and the evaluation of the approach. Section 6 offers a discussion of relevant approaches and draws future work guidelines, while Section 7 concludes the paper.

II. Background

In the domain of requirements engineering, the term traceability is usually defined as the ability to follow the traces (or, in short, to trace) to and from requirements. Two common definitions of requirements traceability are given by Pinheiro [9] as the ability to define, capture, and follow the traces left by requirements on other elements of the software development environment and the traces left by those elements on requirements; and by Gotel and Finkelstein as the ability to describe and follow the life of a requirement in forward and backward directions (i.e., from inception, through specification and development, to subsequent deployment, in addition to on-going refinement and iterations in any of the phases).

The Radio Technical Commission for Aeronautics updated the guidance document DO-178C [10] "Software Considerations in Airborne Systems and Equipment Certification" to address the safety concerns in new technologies such as model-based and object-oriented technologies. The document defines objectives and design

assurance levels for assuring the quality of the software and for an airborne system to perform its intended function with a level of confidence in safety that complies with airworthiness requirements.

The software verification process in DO-178C defines an activity to verify that the system requirements assigned to software have been developed into high-level requirements that meet those system requirements. In order to support this verification, trace data should be generated that show a link between each single system-level requirement and its propagation to test cases. The relationship between a high-level requirement and a test case is bidirectional allowing to trace in forward and backward directions.

Model-driven testing approach, based on transformation rules, uses a model-transformation technique to map a source model to a target one [11]. Model composition approaches automate the composition of heterogeneous models by relying on matching/merging operators [12]. Model-driven approaches move the focus in development from the third-generation programming language coding to more abstract models. This aims to increase productivity and reduce time to market by enabling the use of development concepts closer to the problem domain than those programming languages offer. The main challenge of model-driven development is transforming the high-level models to platform-specific models such that tools can use them for code generation [13]. It is possible to use models horizontally to describe different system aspects; however, they are useful for vertical representation to refine abstractions from the higher to the lower levels, where at the lowest level models use mechanisms based on implementation technology. Significant efforts are needed to work with multiple interrelated models to ensure their overall consistency. Furthermore, using these models can significantly reduce the burden of several other activities like reverse engineering, view generation, application of patterns, and refactoring through automation that is facilitated by the models. Such activities are usually performed as automated processes using one or more source models as input and producing one or more target models, while following a well-defined set of transformation rules. This process is referred to as model transformation.

The guidance document DO-178A [14] introduced at the beginning of 1985 a new technique that supports test coverage and traceability between requirements and tests. This technique, known as requirements-based testing, has been applied in the testing of complex software systems and demonstrated that the systems meet the requirements.

There are several modeling languages to express system requirements as scenarios and numerous languages that can be used to write test scripts. This paper refers to three different notations to capture functional requirements, describes the software description as test specification, and implements and executes scripts against the system under test (SUT). The key points are: (1) system behavioral requirements are formalized and modeled into scenarios representing the same requirements in an alternate Use Case Map (UCM) representation [15], [16], [17], and [18]; the UCM scenarios can be grouped by functionality into sets, for ease of comprehension and maintenance; (2) those UCM models are transformed to abstract test cases using the Test Description Language (TDL) [19], [20] , this process can be viewed as stepwise refinement and model transformation; (3) the obtained TDL abstract scenarios are used as the basis to generate executable test cases in Testing and Test Control Notation (TTCN-3) language. TTCN-3 [21] is a standard language for test specification that is widespread and well-established.

III. Related Work

It is important to establish and maintain relationships among software artifacts because these relationships are useful for many different software engineering activities like software change impact analysis and software validation, verification and testing processes. For instance, the traces can be used to keep models consistent and to identify pairs of related artifacts. These pairs can then be verified and validated against each other. A commonality between MBT and traceability is essential to manage the relationships among different artifacts. Relationship management should assist conception, persistence, and preservation of meaningful relationships across software artifacts in addition to assisting in the destruction of relationships.

Automated MBT approaches exploit two types of relationships: (1) implicit relationships embedded in the tool's algorithms and models, and (2) explicit relationships created and made explicit either automatically by the tool, or manually by the users.

Some approaches as in [22], [23] and [24] use implicit relationships to support test generation, execution and evaluation; while others like in [25] use implicit relationships to support regression testing. Further approaches use explicit relationships to support test generation [26], test execution and evaluation [27], or coverage analysis.

Naslavsky et al. [28] use one kind of behavioral UML model for test generation. A control-flow representation is used along with domain analysis of the parameters of the sequence diagram.

Basanieri et al. [29] use a tool (COW_SUITE) that loads UML models to create explicit relationships as edges in hierarchical trees among them.

Anquetil et al. [30] addressed some of the challenges in developing software product lines in two steps; (1) develop a model-driven framework to identify traceability of variability and (2) specify a metamodel for recording the traceability links.

In [31], the authors integrated a model-driven approach that exploits traceability relationships between monitoring data and architectural model to derive recommended refactoring solutions for the system performance improvement.

Bünder et al. introduce a domain-specific language called Traceability Analysis Language [32] to create and maintain relations of all artifacts that specify, implement, test, or document a software system. The relations are recorded in a

traceability information model and later aggregated to support software development and project management activities with a real-time overview of the state of development.

In [33], the authors adopt the tool (AGEDIS) that uses user-created explicit relationships to execute and evaluate the test scripts. The created relationships map abstract stimuli to method invocations and abstract observations to value checking. In addition, this tool expresses relationships between abstract test suites and test trace results during test execution. Manual coverage analysis is supported via the visualization of the test traces and the abstract test suite that generated them.

In [34], the (AsmL) tool uses user-generated explicit relationships to execute and evaluate abstract test scripts. The use of relationships in the AsmL tool supports the parallel execution of the model and its implementation by relating them and comparing their states.

An approach presented by Abbors et al. [35] provides requirements traceability across an MBT process and the tools used. Additional earlier research addressed using requirement-based testing to support traceability between the requirements and the related testing cases.

Arnold et al. propose a scenario-driven approach [36] (supporting both functional requirements and non-functional requirements) that helps create the traceability between generated and executed test cases, and the executions of an implementation under test.

Furthermore, a model-driven approach combining the strengths of both scenario-based and state-based modeling styles is described in [37]. The tool proposed enables tracing from requirements to testing and from testing to requirements in a round-trip engineering approach.

Pfaller et al. suggest [38] using varying levels of abstraction in development to derive test cases and link them to the corresponding user requirements.

Another approach suggested by Boulanger and Dao [39], where requirement engineering is performed in different phases of the V-model to enable requirements validation and traceability.

Felderer et al., however, focus on model–driven testing of service-oriented systems in a test–driven way [40]. They suggest that the Telling TestStories tool can support traceability among all types of modeling and system artifacts. Marelly et al. discuss linking requirements and testing through the extension of sequence charts with symbolic instances and symbolic variables [41].

IV. Traceability Approach

This work builds on some of the techniques described earlier to create the traceability approach of MBT artifacts. The Ecore trace model is integrated into Eclipse Modeling Framework (EMF) and it is independent of the models it connects. The traceability approach in Fig. 1 [42] shows how system requirements, represented in an abstract model, are propagated through model transformation to more refined models. Furthermore, the traceability approach shows how the relationships among the generated models are created and recorded in a trace model. The first step in the approach is to represent the functional requirements of a system. The use of the modeling tool jUCMNav [43] help describe the system requirements as scenario models in UCM notation. In step 2, the behavioral models, described in step 1, are flattened to scenario definitions using the path traversal algorithm in the jUCMNav tool. Each flattened scenario is transformed, based on transformation rules, to test description in TDL. During the transformation process, the traceability information between the two models (UCM and TDL) are explicitly defined as a trace model. Lastly, test cases generation starts in step 3; it uses the transformed TDL test description models and data model (additional information) to generate the TTCN-3 test cases. Once more, during the process of generating test cases, the traceability information between TDL and TTCN-3 artifacts are explicitly defined and made persistent based on and guided by a traceability scheme.

The key points of the traceability approach are: (1) natural language requirements are described as scenario models in UCM; (2) the UCM models are transformed to test scenario in TDL; and (3) the resulting TDL test scenarios are used along with data model, detailing test data, to generate test cases in TTCN-3. Since the UCM models emphasize behavior and abstract from concrete data, this work focuses on developing a metamodel to support the test data. The developed data model is based on test requirements consisting of three metamodel elements: (1) the UCM responsibilities for message exchange, (2) A set of typed TDL data, and (3) a detailed TTCN-3 data with concrete value. During model transformation, traceability information is defined explicitly into a trace model (tracemodel.ecore). In the following subsections, an example is used to show how relationships among the testing artifacts are created and captured in the trace model during model transformation. The applicability and the evaluation of the approach is demonstrated via case studies in Section 5.

A. Scenarios in UCM Metamodel

The user requirement notation standard suggested UCM notation to capture the functional requirements of a system in terms of visual use case. This latter represents the behavior of a system as a casual scenario composed of responsibilities that can be attached to abstract components. The scenario models, as shown in Fig. 1 (step 1), represent the functional requirements of a system. The UCM models help design and understand systems. The UCM models could be used as a base to derive the test specification cases which in their turn used to develop the test cases.

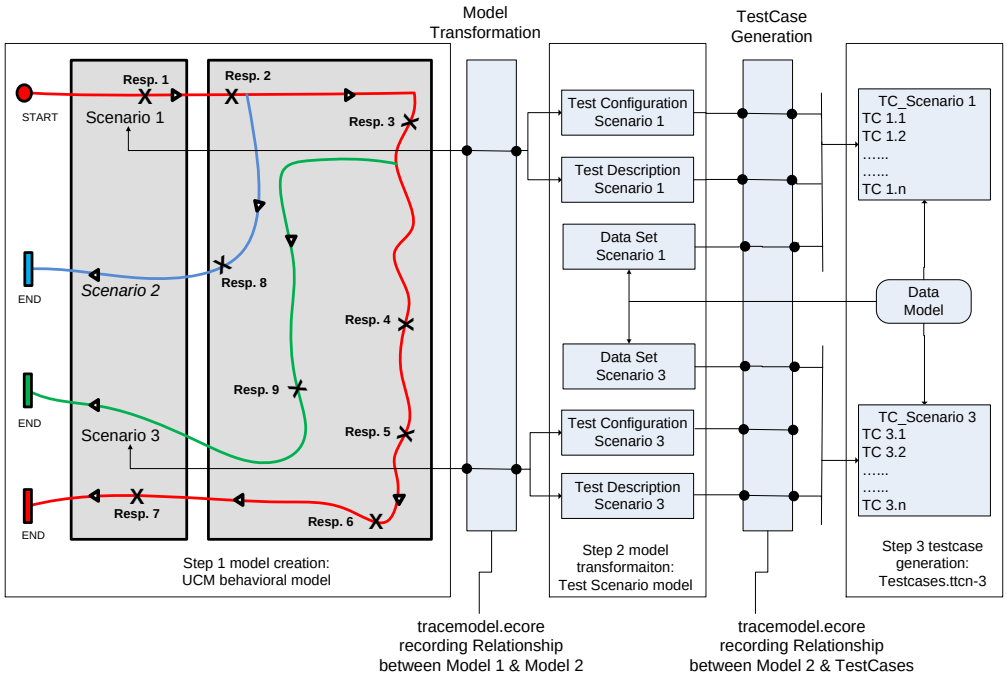


Fig. 1. Traceability Approach Overview.

B. Test Scenarios in TDL Metamodel

The European Telecommunications Standards Institute proposed TDL [44] as a standardized scenario-based approach to specify software test cases as scenarios. TDL is a new standard developed for specifying "formally defined Test Descriptions used for test automation. It offers a high level of abstraction for specifying scenarios beyond programming or scripting languages. TDL can also be used to represent tests generated from other sources like simulators, test case generators, or earlier runs' logs". As described in [45], TDL is a general formal language for representing Test Descriptions which are used mainly for communication between stakeholders as the basis for implementing concrete tests. The TDL design is centered on three separate concepts: (1) the metamodel principle that expresses its abstract syntax; (2) concrete syntax, which is user defined for different application domains; and (3) the TDL semantics that can be found in metamodel elements.

Our approach's main goal is to discover relationships between testing artifacts to support requirement coverage and test evaluation. The model-based test scenario method will support scenario derivation from the UCM behavioral models, and link the relationships from the behavioral model to the test cases. TDL metamodel is used to support the description of scenarios. An instance of TDL metamodel can describe the essential elements of a test scenario such as messages, behavior, actions, interacting components, etc. The TDL test description metamodel, shown in Fig. 2, describes test description based on the exchanged communications between an SUT and a tester.

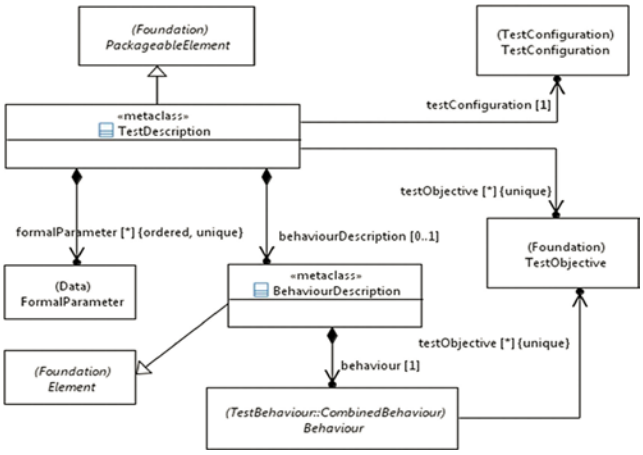


Fig. 2. TDL Test Description Metamodel.

C. Linking UCM Scenarios to TDL Specification

The UCM scenario model shown in Fig. 3 describes the Internet's Domain Name System (DNS) example that verifies whether a DNS server can correctly map a host name to its equivalent IP address.

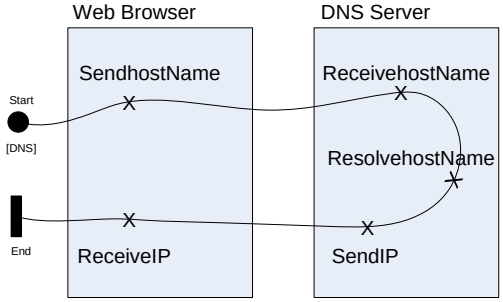


Fig. 3. DNS Scenario Model.

The DNS scenario model has one map contains: a Causal path represented by a wiggly line, two rectangular boxes that represent components Web Browser (Tester) and DNS Server (SUT) and four responsibilities bound to components along the path, and one scenario. The responsibilities elements in UCM are abstract and can represent actions or tasks to be performed by the components. The components themselves are also abstract and can represent software entities (objects, processes, network entities, etc.) as well as non-software entities (e.g. users, actors, processors).

As depicted in Fig. 4, a process (ATC Builder) has been developed to transform the UCM scenario model and data model (additional information) into an abstract test case expressed as a valid TDL.

The outcome of this process is a TDL specification composed of four elements; (1) Data Set, (2) Test Objective, (3) Test Configuration, and (4) Test Description. The DNS scenario model shown in Fig. 3 is transformed into a TDL specification as depicted in Fig. 5.

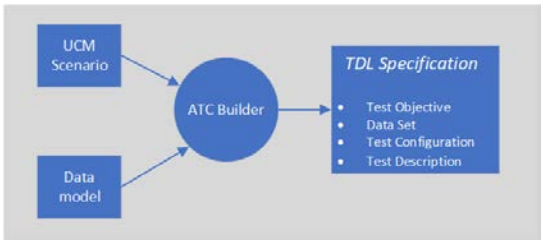


Fig. 4. The Process to Build a TDL Test Specification.

The Component objects, Web browser and DNS server objects, in DNS are transformed into Test configuration items including for example Component Instances, Gate Instance, and Connection. Component Instances can be a part of a Tester or a part of an SUT. Component Instances are connected via the Gate Instance for the exchange of information. The responsibility objects in the DNS scenario model are transformed to Test Description elements such as Action Reference and Interaction. The action to be performed on the Component Instance has an attribute to identify the latter. The gates are used to exchange abstract information which is referenced by the Interaction elements in TDL. This Interaction element could be seen as an exchanged message between source and target.

D. Linking TDL Scenarios to TTCN-3 Test Cases

The UCM scenarios are used as a base to derive the TDL elements. However, the transformed TDL test specification is an abstraction that cannot be executed on SUT. The TDL elements such as Data Instances and Interactions lack concrete details about how to communicate with the SUT. In order for a test case to be executable, it should contain detailed test data and interface specifications. The test inputs for the test cases were developed in a data model during the test analysis and design process. In a UCM scenario, the responsibility object represents an interaction or an action to perform. Therefore, the interaction messages are developed from those responsibilities of nature stimulus/response, mapped into TDL Data Instances, and in turn are developed into a TTCN-3 Template as shown in Table I.

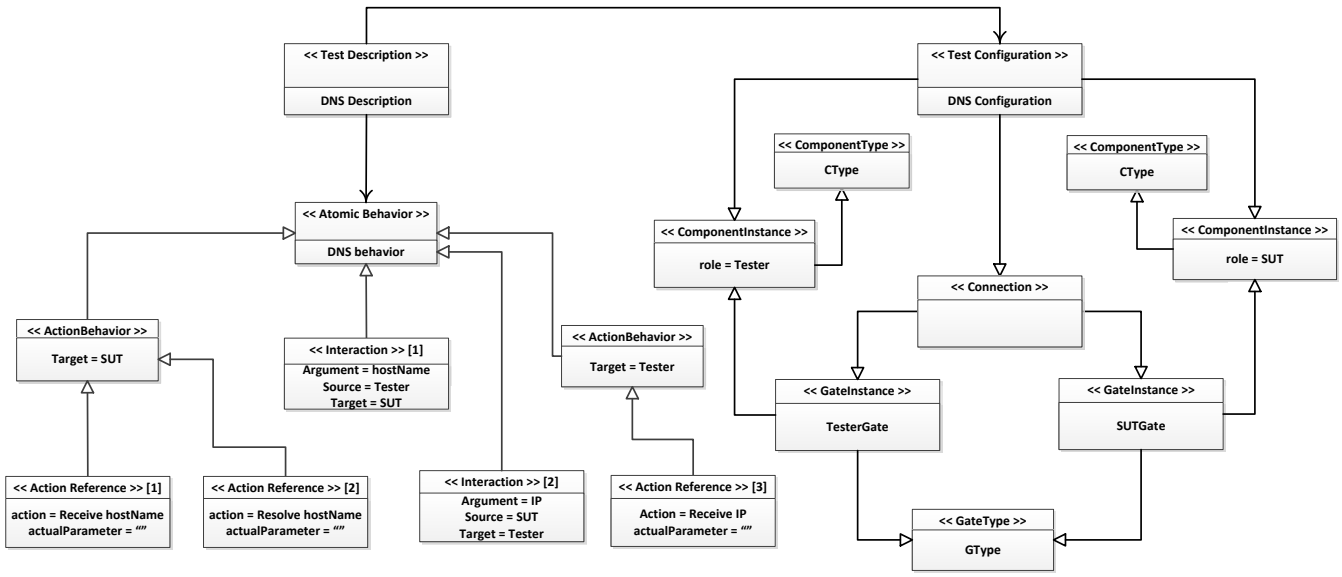


Fig. 5. TDL Metamodel for Test Specification Model.

TABLE I. REFINEMENT OF TEST DATA FROM ABSTRACTION TO CONCRETE [42]

Test Data Input/ Output	**Abstract Data in UCM**	**Data Instance in TDL**	**Data template In TTCN-3**
Stimulus	***SendhostName***	**instance SendhostName**	**Template String** SendhostName :="myHostName"
Response	***ReceiveIP***	**instance ReceiveIP**	**Template String** Receiveip:= "192.124.35.5"
Stimulus	***SendhostName***	**instance SendhostName**	**Template String** SendhostName :="myHostName"
Response	***ReceiveIP***	**instance ReceiveIP**	**Template String** Receiveip:= "192.124.35.5"

Based on data specifications, this work included developing a data model composed of different test data abstraction:

- Stimulus/response: a subset of abstract test data requirements characterized as input and output messages expressed as responsibility objects in UCM.
- Test data instances: the abstract subset of test data requirements is developed to Data Instances and Data Sets in TDL.
- Test data template: using the TTCN-3 templates that define the concrete data, the Data Sets are finally developed and detailed.

The generation of TTCN-3 test cases from the TDL test specification and the data model becomes feasible after applying the transformation rules between the two languages. Transformation rules are defined between TDL and TTCN-3 metamodels resulting in four TTCN-3 modules that together constitute an executable test case: (1) the Configuration module which usually contains several linked test components with unique communication ports, (2) the Description module that consists of behavioral program statements specifying the dynamic behavior of the test components, (3) the Oracle module that contains the expected responses, and (4) the Input module that contains test input data to be transmitted over the communication port. The modules (3) & (4) are derived from the Data Sets and data model. Each requirement to be tested in the data model has an input domain that is subdivided into a set of templates (partitions) and used as a concrete test data. This type of structure will create dependency relationships between a requirement and the relevant test case data. This will help improve regression testing as mentioned in [46]. Since the model transformation starts with flattening the scenario model into scenario definitions, a scenario coverage strategy is applied. Each flattened scenario is transformed to a test scenario and enriched with test data to derive the test cases. This way, straightforward relationships are established between the scenario and the test cases.

E. Traceability Metamodel

In the context of model-driven development, traceability schemes are usually explicitly expressed in metamodels, which are also usually linked to models specifying model transformations. Currently there is no single standardized traceability metamodel. The traces among testing artifacts can be produced on-line, where case traces are stored automatically by a tool as a by-product of the development activity. It can also be done off-line, where traces are recorded (automatically or manually) after the actual development activity has ended. The approach proposed earlier uses a trace metamodel inspired from Jouault et al. [47] that supports traceability. This work's contribution is externalizing and maintaining the relationships between the test-artifact models (i.e. the UCM scenario models, Test scenario models and Test cases models) and recording them in the new trace model. The relationships are recorded semi automatically in the trace model to support various activities like results evaluation, regression testing and coverage analysis. The traceability metamodel to hold the relationships among testing artifacts is defined in UML class relationship diagram as shown in Fig. 6. A class relationship diagram describes the types of objects in the model and selected relationship among them. The relationships can be of type (1) 'Generalization' that relates a specific classifier to a more general classifier. Generalization is denoted by an arrow with an unfilled, triangle head. (2) 'Association' that denotes responsibilities and are shown as lines connecting classes. (3) 'Dependency' where a class A depends on another class B. Dependency is indicated by a dashed line ending at a navigability arrow head. (4) 'Aggregation' can be read as "is part of" or, in the opposite direction as "has a". Aggregation is denoted by an arrowhead drawn as an unfilled diamond. (5) "Composition" implies that the "lifetime" of the parts is bound to the lifetime of the whole. Composition is denoted by an arrowhead drawn as a filled-in diamond.

F. Traceability Scheme

The first step of model creation constructs the UCM model with integrated features (path traversal algorithm) capable of exporting scenario models that conform to the EMF metamodel, Ecore, and implementation of the UCM notations. The implementation of the second step, model transformation, is based on the "UCM scenario to test scenario" model transformation. To support traceability, the transformation tool is extended in this work to create traces that relate the model elements between UCM scenarios and TDL specification. Guided by the traceability scheme defined in Table II, the produced traces in the traceability model called "tracemodel.ecore" were recorded. Implementation of the third step, test case generation and traceability information, takes place when the transformed TDL specifications and the data model developed earlier are ready. These traces were again recorded as a product of the transformation, with the guidance of the traceability scheme.

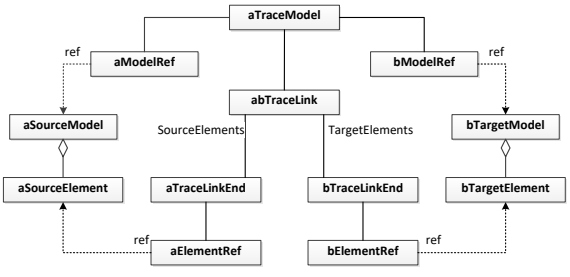


Fig. 6. Traceability Model (Kesserwan Dissertation [42]).

TABLE II. TRACEABILITY SCHEME

Testing artifacts/ Traces	**What information to record**	**Constraints**	**Source**
UCM Scenario	Component, Interaction, Action Reference		Scenario Definition
TDL Test Specification	Test Configuration, Test Description, Gate, Interaction, Action Reference, Data Instance, and Data Set	No duplication in Gate	Connected components
			Set of Interaction
		No duplication in Data Set	Action reference
			Component Interaction
			Data model
TTCN-3 Testcase	Port, Record, Record field, Send, Receive Template, and Function	No duplication in Port	Gate
			Interaction
			Data Set
			Data Instance
			Action reference

V. APPROACH APPLICABILITY AND EVALUATION

The application and the evaluation of the traceability framework have been demonstrated by conducting two case studies from the avionics industry. The first case study is called the landing gear system [48], used to demonstrate the applicability of the approach, where the second one is the FMS used for the evaluation.

A. Test Cases and Trace Model Generation

The description of the landing gear behaviour is captured in UCM scenarios and explained in the following. The goal of the landing gear in an aircraft is to provide support during taxi, take-off and landing. Before landing, the landing order of an airplane is: unlock the landing gear doors, extend the gears and lock the landing gear doors. Fig. 7 depicts a successful deployment of extending sequence scenario [DeploymentSucceeded], and two unsuccessful deployment scenarios; [DeploymentFailed] and [NormalModeFailed].

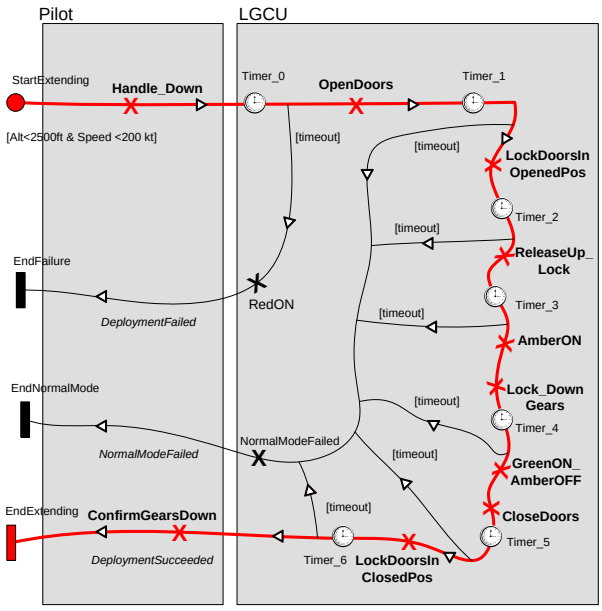


Fig. 7. Visual UCM Scenario Describing the Extending Sequence Case.

The creation of the UCM model was described as step 1 of the approach (Fig. 1). The next step is to transform the UCM model into a TDL test specification, and create the traceability information. The test data for the successful scenario [DeploymentSucceeded] is shown in Table III.

The graphical representation of the transformed model, composed of test description and test configuration elements, is depicted in Fig. 8. Traceability information for the test configuration is depicted in Fig. 9, while part of the traceability information for the test description is depicted in Fig. 10.

In Fig. 9, the traceability model is named TraceUCMModel2TDLModel. It relates models UCMScenarioModel and TDLTestScenarios. It has one trace link named DSScenarioTraceLink that relates the UCMDSScenario in the UCMScenarioModel to the TDLDSTTestSpecification in the TDLTestScenarios. DSSScenarioTraceLink has many children; the figure shows the link DSTestConfigurationTraceLink, which relates the component Instances (Pilot and LGCU) in the UCMDSScenario to the gate instances (Tester and SUT) in the TDLDSTestSpecification.

In Fig. 10, the trace link DSSScenarioTraceLink has another child DSTestDescriptionTraceLink, which relates the interactions and action references in the UCMDSScenario to the interactions and action references in the TDLDSTestSpecification. The figure shows one “Interaction” and one “Action Reference”.

The last step in the approach is the generation of test cases and the creation of the traceability information in the TDL test model and the generated test cases. Information from the data model in Table II, from the trace model in Fig. 10 and from the test specification model in Fig. 8 is used to complete the step. The data model is developed from the testing requirement and represents the input space for the scenario model [DeploymentSucceeded] under transformation. The instances in the data model are grouped into two sets; stimulus (Tester) and response (SUT) to build the TDL Data Sets elements. Each Data Set is mapped to records and variables elements in TTCN-3 using the transformation rules between the two languages. In Fig. 11, the trace link DSSScenarioTraceLink has a child DSTestDataModuleTraceLink, which relates the Data Set, Data Instance and Interaction in the TDLDSTestSpecification to the Record, Record field and Send in the TC_DS_[seq]. The figure shows one “Data Set”, one “Instance” and one Interaction. The TDL test scenario [DeploymentSucceeded] is transformed into a test case in TTCN-3. The approach defined in [8] applies structural transformation where each TDL element is transformed into a number of TTCN-3 modules. Based on transformation rules, the resulting test case is composed of three types of modules: (1) a Test Configuration module, (2) a Test Description module, (3) and a Data module. The TTCN-3 data module is refined with test input and expected output when this data becomes available. A new test case is added “TC_DS_01” to the test suite “TTCN-3_DC_TestSuite” for each new pair of test input and expected output found in the Data model in Table II.

TABLE III. THE DEVELOPMENT OF TEST DATA FOR [DEPLOYMENTSUCCEEDED] SCENARIO [42]

Test Data Requirement	UCM responsibility Stimulus/Response	TDL Data Instance	TTCN-3 Template
Send stimulus when handle is pushed down	***Handle_Down***	**instance Handle_Down**	**Template String** Handle_Down_Type
Receive a response when locking doors in opened position	***LockDoorsInOpenedPos***	**instance LockDoorsInOpenedPos**	**Template String** LockDoorsInOpenedPos_Type
Receive a response when Gear is in transition	***AmberON***	**instance AmberON**	**Template String** AmberON_Type
Receive a response when locking Gears in down position	***GreenON_AmberOFF***	**instance GreenON_AmberOFF**	**Template String** GreenON_AmberOFF_Type
Receive a response when locking doors in closed position	***LockDoorsInClosedPos***	**instance LockDoorsInClosedPos**	**Template String** LockDoorsInClosedPos_Type

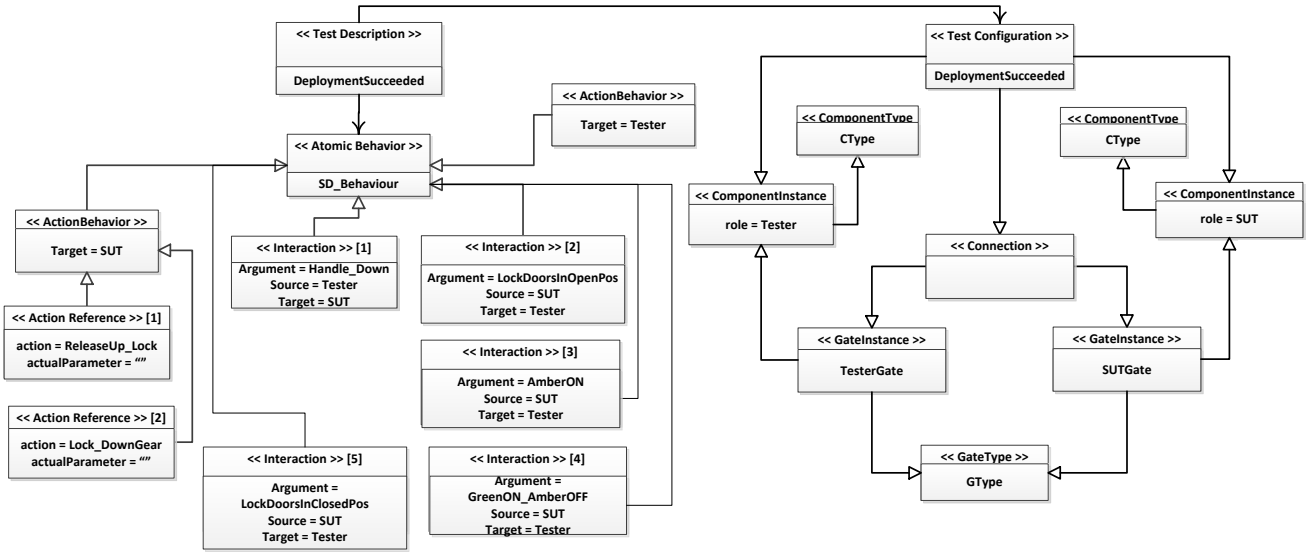


Fig. 8. Test Specification Model for [DeploymentSucceeded] Scenario (Kesserwan Dissertation [42]).

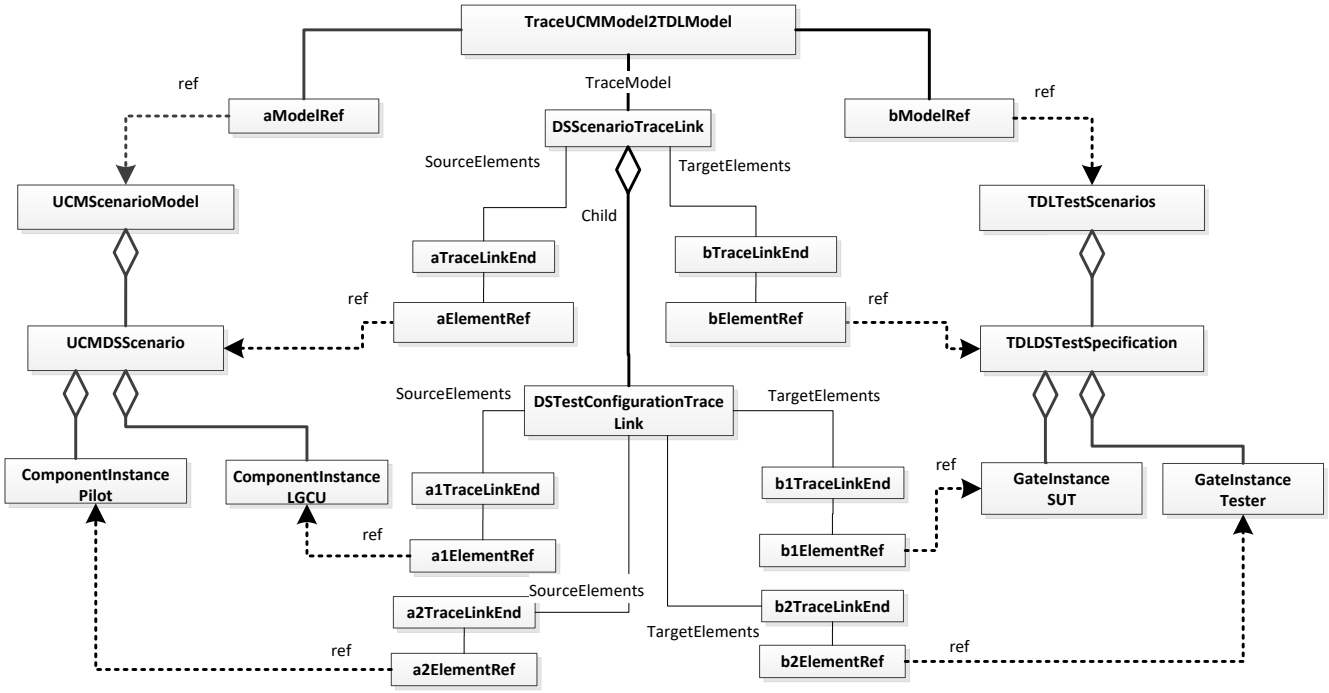


Fig. 9. Traceability Model shows Traceability Links between the UCM and TDL Models for [*DeploymentSucceeded*] Scenario (Kesserwan Dissertation [42]).

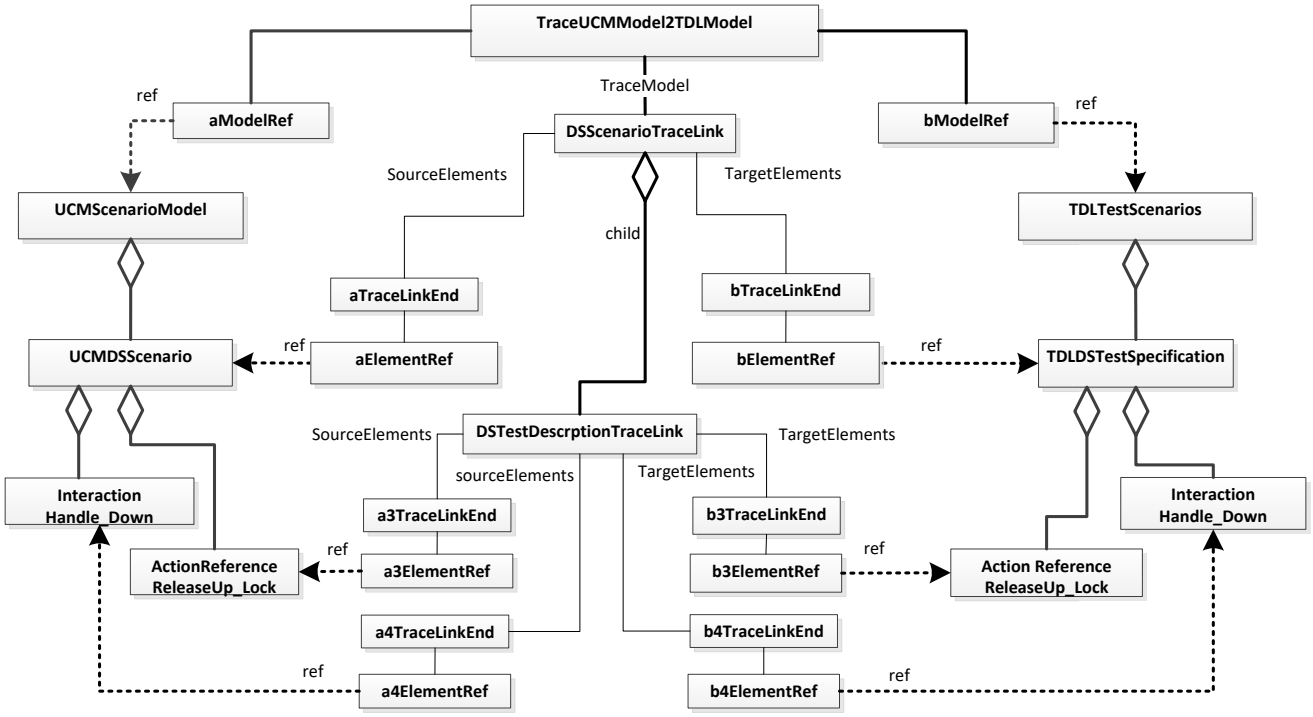


Fig. 10. A Small Part of Traceability Links between the Two Models for [DeploymentSucceeded] Scenario (Kesserwan Dissertation [42]).

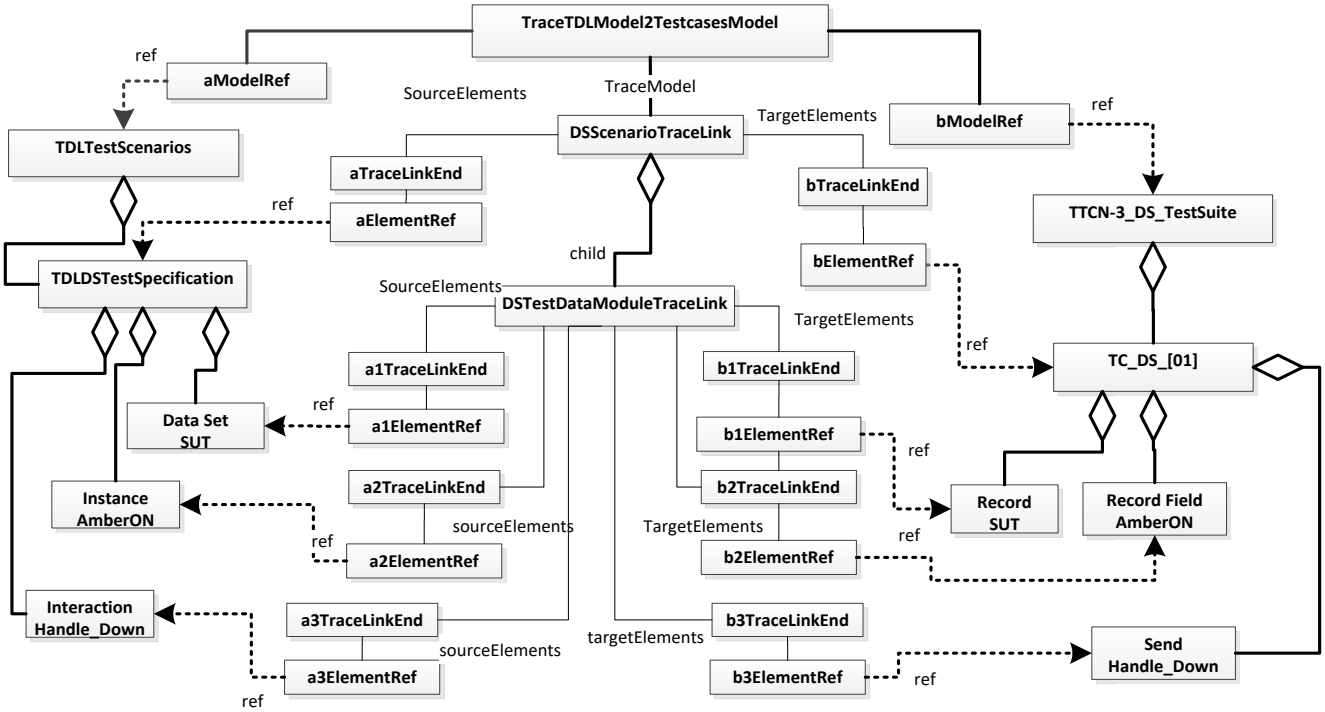


Fig. 11. Traceability Information between TDL and TTCN-3 (Kesserwan Dissertation [42]).

B. Traceability Links and Alignment with Test Result

To evaluate the extended testing methodology in this work, the experiment method described in [8] is reused to generate the test case. The new obtained result is a trace model (tracemodel.ecore) which relates UCM scenario models to TTCN-3 test cases grouped in test suites. Each test case, generated with a unique identifier, is a sequence of actions and interactions with defined input parameter values and output parameter values. The execution of the test case results in the assignment of a test verdict; pass or fail. In the trace model, the links between requirements and test cases may have several possible cardinalities:

- One-to-one: one requirement is tested exactly by one test case and this test case tests only this requirement.
- One-to-many: one requirement is tested by several test cases and these test cases participate to test only this requirement.
- Many-to-many: one requirement is tested by several test cases, which are used to test several requirements.

Fig. 12 shows the relationships between the testing artifacts for the [DeploymentSucceeded] scenario. The traceability link DSScenarioTraceLink[1] relates the model UCMDSScenario to the model TDLDSTestSpecification which is related to several test cases via the traceability link DSScenarioTraceLink[2]. The generated test cases are children of the test suite TTCN-3_DS_TestSuite.

The trace model takes a significant importance in the test generation process. On one hand, it provides a clear meaning for each generated test case: the tested requirement(s) gives the purpose of the associated test case(s). It is a kind of rationale for the generated test suite. On the other hand, the trace model exhibits clearly which requirements are actually tested (and how), and which requirements are not tested. For the not tested requirements, this suggests completing the test suite to obtain full functional coverage. During test execution of the test case, the traceability links in the trace model help to identify the related requirements when it fails. Similarly, when the test case passes, they certify that the related requirements were implemented and tested.

C. Requirement Coverage and Compliance with DO-178C

The trace model helped analyze the generated TDL test description from UCM models to check if the test cases cover the requirements. The trace model showed full coverage between UCM scenarios and their developed TDL specifications. The trace model realized complete requirement and scenario coverage. For each path in the UCM model, there is a TDL test scenario linked to it and the number of links in the trace model equals the number of scenarios found in the UCM model.

Furthermore, the trace model helped analyze the generated TDL test description to check if they are actually traceable to the original software requirements (UCM elements). The trace model meets the traceability objective as defined by DO-178C standard where an association between a requirement and its related items is necessary. The trace model contains links between the UCM models and the TDL test scenarios which in turn are traced to the generated test cases in TTCN-3. Therefore, compliance with DO-178C is achieved.

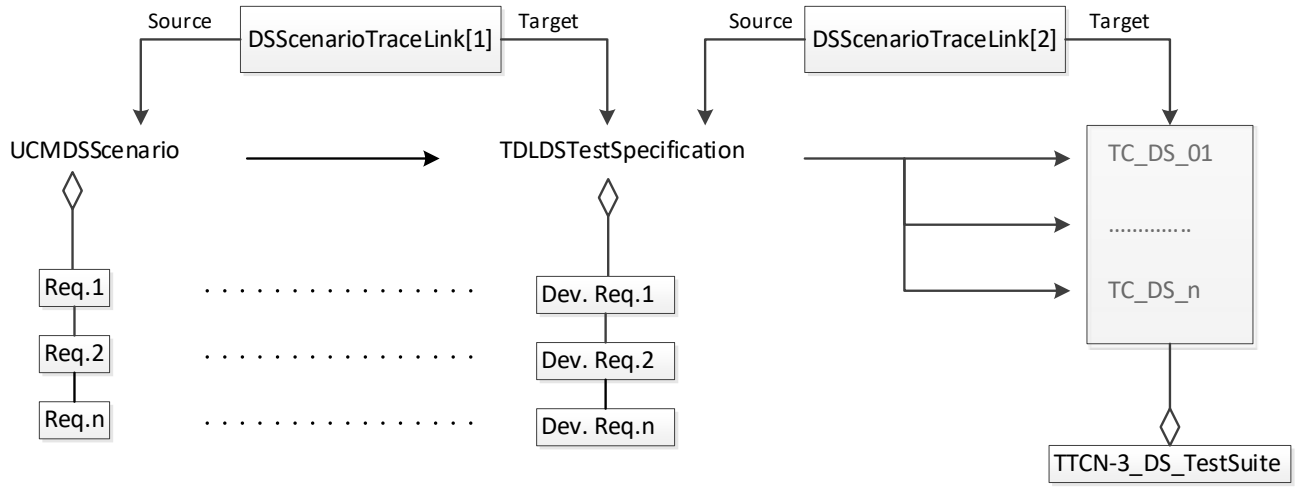


Fig. 12. Traceability Links among Testing Models for [DeploymentSucceeded] Scenario.

VI. Discussion and Future Work

Similar to the approaches discussed in the related work section (Section 3), this paper proposes to create traceability links among testing artifacts. However, this work differs from them as the proposed method extends the model-driven testing methodology to create explicit relationships in a trace model among testing artifacts. The approach creates UCM behavioral models and relates them to test cases via abstract test models during model transformation where n-ary links among models could be visualized. This is an important factor in visualizing relationships among models because it is almost impossible to represent more than one link in a two-dimensional traceability matrix in an understandable way. Moreover, the number of relationships in traceability matrices is high and fixed. The trace model records a small number of relationships from model to a testcase to enable the support for model-based coverage analysis, visualizing traceability and result evaluation.

Another important difference is creating a semiautomatic process for trace recording. This reduces some of the repetitive and time consuming tasks testers need to do to generate these traceability connections. Most models discussed require manual recording. This also distinguishes this work from the earlier research in this specific topic as it extends the scope and capabilities of the model developed and improves its processes.

This work is the start of research efforts to offer more effective ways to ensure traceability and create better pathways for validation. Following this contribution, future work will focus on enhancing the model to provide additional traceability aspects and addressing some of the current limitations. More research into enhancing the traceability process such that it could use additional sources (other than UCM) to provide access to non-functional requirements. This will further improve the traceability model and provide a more robust coverage of requirements. In addition, methods to automate the steps in this process will be investigated and a fully automated process of recording traces in the trace model will be explored. This will create a faster and more effective process for test traceability.

As a result, non-functional requirements, generally not captured by UCM, cannot be used. In addition, the semi-automatic recording improved the process, yet it still requires manual work to complete the process.

VII. Conclusion

Our main contribution in this paper is the proposal and presentation of a model-based approach that leverages available methods to generate test artefacts based on model transformations. This approach enables creating traceability links among testing artifacts. It also extends the transformation methodology to create and document relationships as a set of metadata in a trace model through consecutive transformation steps. A traceability scheme with constraints that determines which testing artifacts and at which level of detail the traces can be recorded was defined. The proposed traceability scheme guides the recording of traces (manual) and makes them persistent. Relationships are created and made explicit among scenario definitions in UCM models, their test specifications in TDL notation, and the corresponding test suite scenario in TTCN-3 language. The documented relationships in a trace model enable the support for visualizing traceability, coverage analysis and test result evaluation. This paper shows the developed infrastructure and workflow for MBT that applies model transformation and test generation techniques to create test scenarios, test cases, and traceability models.

References

[1] Tanvir Hussain and Robert Eschbach, "Automated Fault Tree Generation and Risk-Based Testing of Networked Automation Systems," in Proceedings of 15th IEEE Conference on Emerging Technologies and Factory Automation (ETFA 10) Bilbao, Spain, 2010.

[2] Lago, P., Muccini, H., van Vliet, H.: A scoped approach to traceability management. J. Syst. Softw. 82(1), 168–182 (2009).

[3] Winkler, Stefan, and Jens von Pilgrim. "A survey of traceability in requirements engineering and model-driven development." Software & Systems Modeling 9.4 (2010): 529-565.

[4] Galvao, I., & Goknil, A. (2007, October). Survey of traceability approaches in model-driven engineering. In 11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007) (pp. 313-313). IEEE.

[5] Aichernig, Bernhard K., Wojciech Mostowski, Mohammad Reza Mousavi, Martin Tappler, and Masoumeh Taromirad. "Model learning and model-based testing." In Machine Learning for Dynamic Software Analysis: Potentials and Limits, pp. 74-100. Springer, Cham, 2018.

[6] J. Dick, Faivre, A., Automating the Generation and Sequencing of Test Cases from Model-Based Specifications, Springer-Verlag, 1993.

[7] D. J. Richardson, Aha, S. L., O'Malley, T. O., Specification-based test oracles for reactive systems, Proceedings of the 14th international conference on Software engineering, ACM Press, Melbourne, Australia, 1992, pp. 105-118.

[8] Kesserwan, N., Dssouli, R., Bentahar, J., Stepien, B. and Labrèche, P., 2017. From use case maps to executable test procedures: a scenario-based approach. Software & Systems Modeling, pp.1-28.

[9] Pinheiro, F.A.C.: Requirements traceability. In: Sampaio do Prado Leite, J.C., Doorn, J.H. (eds.) Perspectives on Software Requirements, pp. 93–113. Springer, Berlin (2003).

[10] DO-178C, available from RTCA at www.rtca.org. Retrieved 01/22/2021.

[11] Bernhard Schatz. 2011. 10 Years Model-Driven -- What Did We Achieve?. In Proceedings of the 2011 Second Eastern European Regional Conference on the Engineering of Computer Based Systems (ECBS-EERC '11). IEEE Computer Society, Washington, DC, USA, 1-. DOI=http://dx.doi.org/10.1109/ECBS-EERC.2011.42.

[12] Kienzle, Jörg, et al. "A unifying framework for homogeneous model composition." Software & Systems Modeling 18.5 (2019): 3005-3023.

[13] Eclipse Modeling Framework (EMF), available at http://www.eclipse.org/modeling/emf/, retrieved 01/22/2021.

[14] DO-178A Software Considerations in Airborne Systems and Equipment Certification, Document Number: DO-178A, Issue Date: 3/22/1985, Committee: SC-152, Category: Software.

[15] Zaman, Qamar uz, Aamer Nadeem, and Muddassar Azam Sindhu. "Formalizing the use case model: A model-based approach." Plos one 15, no. 4 (2020): e0231534 Buhr, R.J.A.: Use case maps as architectural entities for complex systems. IEEE Trans. Softw. Eng. 24(12), 1131–1155 (1998).

[16] Buhr, R.J.A.: Use case maps as architectural entities for complex systems. IEEE Trans. Softw. Eng. 24(12), 1131–1155 (1998).

[17] ITU-T Z.151 - the International Telecommunication Union, available at https://www.itu.int/en/pages/default.aspx, retrieved 01/22/2021.

[18] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.9896&rep=rep1&type=pdf.

[19] G. Spanoudakis, Zisman, A., Software Traceability: A Roadmap, Advances in Software Engineering and Knowledge Engineering, World Scientific Publishing, 2005.

[20] Philip Makedonski, Gusztav Adamis, Martti Käärik, Andreas Ulrich, Marc-Florian Wendland, Anthony Wiles. “Bringing TDL to users: A Hands-on Tutorial” User Conference on Advanced Automated Testing (UCAAT 2014), Munich.

[21] TTCN-3 Standards, available at http://www.ttcn-3.org/index.php/downloads/standards.

[22] F. Fraikin, Leonhardt, T., SeDiTeC — Testing Based on Sequence Diagrams, 17th IEEE International Conference on Automated Software Engineering, 2002, pp. 261 - 266.

[23] Gagarina, Larisa G., Anton V. Garashchenko, Alexey P. Shiryaev, Alexey R. Fedorov, and Ekaterina G. Dorogova. "An approach to automatic test generation for verification of microprocessor cores." In 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 1490-1491. IEEE, 2018.

[24] J. Wittevrongel, Maurer, F., SCENTOR: Scenario-Based Testing of E-Business Applications, Tenth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2001, pp. 41 - 46.

[25] L. C. Briand, Labiche, Y., A UML-Based Approach to System Testing, 4th International Conference on the Unified Modeling Language (UML), Toronto, Canada, 2001, pp. 194-208.

[26] F. Basanieri, Bertolino, A., Marchetti, E., The Cow_Suite Approach to Planning and Deriving Test Suites in UML Projects, Proceedings of the 5th International Conference on The Unified Modeling Language, Springer-Verlag, 2002, pp. 383-397.

[27] W. Grieskamp, Nachmanson, L., Tillmann, N., Veanes, M., Test Case Generation from AsmL Specifications - Tool Overview, 10th International Workshop on Abstract State Machines, Taormina, Italy, 2003.

[28] Naslavsky, Leila, Hadar Ziv, and Debra J. Richardson. "Towards traceability of model-based testing artifacts." Proceedings of the 3rd international workshop on Advances in model-based testing. ACM, 2007.

[29] F. Basanieri, Bertolino, A., Marchetti, E., The Cow_Suite Approach to Planning and Deriving Test Suites in UML Projects, Proceedings of the 5th International Conference on The Unified Modeling Language, Springer-Verlag, 2002, pp. 383-397.

[30] Anquetil, N., Kulesza, U., Mitschke, R., Moreira, A., Royer, J. C., Rummler, A., & Sousa, A. (2010). A model-driven traceability framework for software product lines. Software & Systems Modeling, 9(4), 427-451.

[31] Arcelli, D., Cortellessa, V., Di Pompeo, D., Eramo, R., & Tucci, M. (2019, March). Exploiting architecture/runtime model-driven traceability for performance improvement. In 2019 IEEE International Conference on Software Architecture (ICSA) (pp. 81-90). IEEE.

[32] Bünder, H., Rieger, C., & Kuchen, H. (2017). A Model-Driven Approach for Evaluating Traceability Information. Model-Driven Software Development, 436.

[33] A. Hartman, Nagin, K., The AGEDIS tools for model based testing, 2004 ACM SIGSOFT international symposium on Software testing and analysis, ACM Press, Boston, Massachusetts, USA, 2004, pp. 129-132.

[34] W. Grieskamp, Nachmanson, L., Tillmann, N., Veanes, M., Test Case Generation from AsmL Specifications - Tool Overview, 10th International Workshop on Abstract State Machines, Taormina, Italy, 2003.

[35] F. Abbors, D. Truscan, and J. Lilius, "Tracing requirements in a model-based testing approach," in 2009 First International Conference on Advances in System Testing and Validation Lifecycle (VALID), Piscataway, NJ, USA, 2009, pp. 123-8.

[36] D. Arnold, J. P. Corriveau, and Shi Wei, "Modeling and validating requirements using executable contracts and scenarios," in 8th ACIS International Conference on Software Engineering Research, Management and Applications (SERA), CA, USA, 2010, pp. 311-20.

[37] A. Goel, B. Sengupta, and A. Roychoudhury, "Footprinter: Round-trip engineering via scenario and state-based models," in 31st International Conference on Software Engineering - Companion Volume - ICSE-Companion, Piscataway, NJ, USA, 2009, pp. 419-420.

[38] C. Pfaller, A. Fleischmann, J. Hartmann, et al., "On the integration of design and test: A model-based approach for embedded systems," in Proceedings of the 2006 international workshop on Automation of software test (AST) 2006, pp. 15-21.

[39] J. L. Boulanger and V. Q. Dao, "Requirements engineering in a model-based methodology for embedded automotive software," in IEEE International Conference on Research, Innovation and Vision for the Future in Computing 484 & Communication Technologies(RIVF), Ho Chi Minh City, Vietnam, 2008, pp. 263-268.

[40] M. Felderer, P. Zech, F. Fiedler, et al., "A Tool based Methodology for System Testing of Service-oriented Systems," in Second International Conference on Advances in System Testing and Validation Lifecycle (VALID), Los Alamitos, CA, USA, 2010, pp. 108-13.

[41] R. Marelly, D. Harel, and H. Kugler, "Multiple instances and symbolic variables in executable sequence charts," in 17th International Conference on Object-Oriented Programming, Systems, Languages and Applications (OOPSLA 2002), USA, 2002, pp. 83-100.

[42] Kesserwan, N. (2020). Automated Testing: Requirements Propagation via Model Transformation in Embedded Software (Doctoral dissertation, Concordia University).

[43] http://istar.rwth-aachen.de/tiki-index.php?page=jUCMNav

[44] ETSI ES 203 119-1 V1.3.1 standard, available at http://www.etsi.org/deliver/etsi_es/203100_203199/20311901/01.03.01_60/es_20311901v010301p.pdf, retrieved 01/22/2021.

[45] Ulrich,A., Jell, S.,Votintseva, A.,Kull, A.: The ETSI TestDescription Language TDL and its application. In: 2014 2nd International Conference on Model-Driven Engineering and Software Development (MODELSWARD), pp. 601–608. IEEE (2014, January).

[46] P. Stocks, Carrington, D., A Framework for Specification-Based Testing, IEEE Transactions on Software Engineering, 1996, pp. 777-793.

[47] M. Didonet Del Fabro, Bézivin, J., Valduriez, P., Weaving Models with the Eclipse AMW plugin, Eclipse Modeling Symposium, Eclipse Summit Europe 2006, Esslingen, Germany, 2006.

[48] Boniol, F.,Wiels, V.: The landing gear system case study. In: ABZ 2014: The Landing Gear Case Study, pp. 1–18. Springer (2014).

Space Mining Robot Prototype for NASA Robotic Mining Competition Utilizing Systems Engineering Principles

Tariq Tashtoush[1]*, Jesus A. Vazquez[2], Julian Herrera[3], Liliana Hernandez[4], Lisa Martinez[5], Michael E. Gutierrez[6], Osiris Escamilla[7], Rosaura E. Martinez[8], Alejandra Diaz[9], Jorge Jimenez[10], Jose Isaac Segura[11], Marcus Martinez[12]

School of Engineering, Texas A&M International University, Laredo, TX, 78041 USA

***Abstract*—The 2017 National Aeronautics & Space Administration (NASA) Robotic Mining Competition (RMC) is an outstanding opportunity for engineering students to implement all the knowledge and experience that they gained in the undergraduate years, in building a robot that will provide an intellectual insight to NASA, to develop innovative robotic excavation concepts. For this competition, multiple universities from all over the U.S. will create teams of students and faculty members to design and build a mining robot that can traverse, mine, excavate at least 10 kg of regolith, then deposit it in a bin in the challenging simulated Martian terrain. Our team's goal is to improve on our current design and overcome DustyTRON 2.0's limitations by analyzing them and implementing new engineering solutions. The process to improve this system will enable our team members to learn mechanical, electrical, and software engineering. DustyTRON 3.0 is divided into three sub-teams, namely, Mechanical, Circuitry, Software sub-teams. The mechanical team focused on solving the mechanical structure, robot mobility, stability, and weight distribution. The circuitry team focused on the electrical components such as batteries, wiring, and motors. The Software team focused on programming the NVidia TK1, Arduino controller, and camera integration. This paper will outline the detailed work following systems engineering principles to complete this project, from research, to design process and robot building compete at the Kennedy Space Center. Only 54 teams were invited to participate from all over the US and DustyTRON team represented the state of Texas and placed the 29th and awarded the "Innovative Design" award.**

Keywords*—NASA robotic mining competition; mining robot; ice regolith; autonomous; NASA; space exploration; systems life-cycle, mechanical structure design, control system, systems engineering; software development

I. Introduction

As a leader in space exploration, the National Aeronautics and Space Administration (NASA) developed several unmanned robots, which were sent to the Moon and Mar in exploration missions to navigate the highly hazardous planets ecosystem and mine the available resources that will be converted to the needed energy (Oxygen and Hydrogen) before sending any human astronauts [1–12]. This technology provided the highest level of human safety and lowered space transportation costs.

The NASA Robotic Mining Competition (RMC) was started to engage university-level engineering students to design, build, operate and compete with a robot that can be sent to space for a Martian chaotic terrain exploration. The off-world mining mission will be simulated where the robot will traverse and excavate simulated resources called regolith (Black Point-1 or BP-1) and ice (gravel), then return and deposit the excavated mass into a collector bin.

The eighth annual NASA Robotic Mining Competition (RMC) took place on May 22-26, 2017 at the Kennedy Space Center. This engineering challenge brought fifty-four U.S. university teams came to compete and show their unique and creative robotics design. DustyTRON Robotic team from Texas A&M International University (TAMIU), fulfilled the competition goals based on NASA guidelines and RMC requirements [13–16]. This work marked our third participation in the RMC competition.

Each robot will have two ten-minutes trials to complete the mission. The field will be a 3.78m x 7.38m arena which will be separated into three sections: starting area, obstacle area with rocks and craters, and a mining area. At the beginning of each trial, the robots were placed in the starting area at random positions and orientations. Then robots must traverse through the obstacle area which will contain two craters and three randomly placed rocks to reach the mining area. Once in the mining area, the robot needs to excavate then return to the starting area where a collection bin will be located to deposit the collected regolith. If time permits the robot will return to collect more regolith from the mining area.

The paper is organized as follows: Section II covers the available literature and NASA explore the space activities, Section III is a system requirements summary, Section IV illustrates all preliminary designs, Section V describes concept operation, Section VI shows the different systems' hierarchy, Section VII details the robot interface, Section VIII is risk management analysis, Section IX is the trade-off analysis, verification of System Meeting Requirements in Section X, Section XI reliability, Section XII summarizes the competition results, and Section XII is the paper conclusion and the future plan.

II. Literature Review

NASA's efforts and Robot exploration have always been around for many years where robots are used to collect data

and to see how their actions and experiences can help us figure out ways to reach and live in space to enhance the future of mankind. Going back to the moon and exploring Mars have always been a goal for the U.S in the past centuries. Lunar mission and deep-space exploration can comply with the Global Exploration Road-map and the National Research Council. This mission name is called ALCIDES. ALCIDES will use some of the previous systems that were used in the HERCALES exploration, such as the Orion module, the Boeing Reusable Lander, the Ariane 6, the Falcon Heavy, the Space Exploration Vehicle, the Space Launch System, and the Evolvable Deep-Space Habitat placed in EML2. Robots and humans will need to work together to meet their goals, autonomously, and cooperate utilizing all the available technologies nowadays.

NASA Robotic Mining Competition (RMC) was stated due to recent NASA missions to Mars' discoveries, robots such as "Curiosity" and orbiting satellites taking pictures and videos showed a large amount of water in form of water ice and hydrated minerals on Mars [13–21]. Water sources formed millions of years ago were determined to be a result of clay and clay-like minerals on the surface or underground of Mars and Moon. Collecting these resources especially water will allow the humans' dream of living off the mainland. These resources can be utilized to provide humans with the required energy for rocket propellants, growing plants and sustaining astronauts, and protecting them in such a harsh environment. These minerals sources must be mined from the surface or buried deep in the ground.

NASA Robotic Mining Competition is a challenge for university-level undergraduate students from all over the United States (US). Students are tasked to design and construct a space-capable robot to traverse simulated Martian terrain and conduct a complete mining mission for the water and minerals. The mining robot must excavate the regolith simulant and/or the ice simulant that is located $30 - 50\ cm$ deep then travel back to the simulated space station collection bin to deposit the collected resources. In addition to the fact that the robot must be space-focused, NASA added few complexities to the challenge such as the robot has to be limited in size and weight, can tolerate the abrasive characteristics of the regolith, can be teleoperated or completely autonomous, and power/bandwidth-efficient.

Students participating in this competition can develop innovative robotic excavation concepts that allow NASA can use such excavation devices for future missions to advance human spaceflight and NASA space exploration operations. More info about this competition can be at https://www.nasa.gov/offices /education/centers/Kennedy/technology/nasarmc.html.

The NASA RMC started in its original format in 2010 as NASA Lunabotics Competition [14]. In 2011, it was open to undergraduate and graduate student teams enrolled in colleges or universities worldwide. But in 2014, due to NASA budgetary constraints, the competition was limited to teams from United States colleges or universities. In 2020, NASA transited to a Lunar-focused competition, and Table I represents the competition year, name, and the allowed countries to participate [6].

TABLE I. NASA ROBOTIC MINING COMPETITION HISTORY [6]

Competition Year and Name	Competition Participants
(2010) Lunabotics	USA
(2011) Lunabotics	USA, Bangladesh, Canada, Colombia, India, Spain
(2012) Lunabotics	USA, Bangladesh, Canada, Colombia, India, Mexico, Romania, South Korea
(2013) Lunabotics	USA, Australia, Bangladesh, Canada, Colombia, India, Mexico, Poland
(2014-2019) RMC	USA
(2020-present) RMC: Lunabotics	USA

Many previously participating teams in NASA RMC presented their robots' design and operation following NASA requirements [16, 22–28].

DustyTRON team utilizes this paper to present the implementation of system engineering concepts and processes in real-life problems and innovative solutions of space mining robots. The team participated previously in NASA RMC where they built mining robots DustyTRON 1.0 (2015) and 2.0 (2016), as shown in Fig. 1 and 2, respectively.

Fig. 1. DustyTRON 1.0 Robot - RMC 2015 [6].

Fig. 2. DustyTRON 2.0 Robot - RMC 2016 [6].

The DustyTRON 3.0 is the improved design of

DustyTRON 2.0, and the team consists of students that have participated before, seniors taking the class, and underclassmen interested in constructing a mining robot. DustyTRON 3.0 has a similar overall mechanical structure as DustyTRON 2.0, but with several improvements. Additionally, this paper includes a detailed analysis of the fully functional mining robot DustyTRON 3.0 to meet certain specifications including size dimension ($1.5mX0.75mX0.75m$), weight ($80Kg$ max), and mechanism (traverse, excavate, and deposit). The team's design theory and Quality Function Deployment (QFD) analysis will be the core of this project. Several designs were developed and evaluated based on multiple criteria such as design to build, mobility, weight, and budget, then followed by decision-making to select one optimum design.

DustyTRON 3.0 was split into three sub-teams: 1) mechanical design and construction, 2) electrical circuitry design, and 3) software development.

- ***Mechanical design and construction team*** focused on a robot structure development where the robot must have a strong structure that moves easily while keeping lightweight, an excavation mechanism, and a regolith collection and deposit mechanism. They will improve the rigidity of the middle structure of the robot, where the excavation mechanism will be mounted and enhance the steering system.
- ***Circuitry team*** will link the mechanical and software components together to achieve a fully functional robot. They will improve on the electrical components and storage for easy accessibility and monitor-ability, and safe from any external influence. Cables will be routed in different layouts so that troubleshooting and repairs will be easier and faster in case of a problem, which will reduce the risk of an electrical short, electrical interference, or electrical failure significantly.
- ***Software development team*** worked on developing the autonomous functionality by moving to System-On-Chip (SoC) and microprocessor system. Intercommunication between the SoC and Microprocessor will be conducted through a serial interface while a secure connection between the robot and the control station will be used. The autonomous mode will utilize OpenCV (Computer Vision) library for image and object detection for Xbox One Kinect and IP cameras.

DustyTRON 3.0 team planned to build a robot based on DustyTRON 2.0 in order to reduce the total budget, by providing improvements and solutions to last year's design problem. The mechanical team's improvements will include motors and steering systems modifications, which was estimated to be $2000. For both circuitry and software teams, the budget estimation was $1000 because last year components will be recycled and used for this year's robot. Table II shows the estimated budget and the actual cost.

The actual budget of DustyTRON 3.0 had been changed along the building process due to sudden failure of electrical components such as motor drives and voltage regulators, hardware parts, such as t-slotted beams deformation, and wheel design changes. With extensive research and some educational discounts and donations, the total cost was lowered. One important note is that the software team did not have to make any significant purchases to prepare for the 2017 NASA RMC competition, as previous years of competing had provided the team with all the physical components to build the software needs. The digital aspect of needed materials required no purchasing since programs like TurboVNC, PuTTY, Arduino IDE, and Ubuntu were all free.

TABLE II. DUSTYTRON 3.0 ESTIMATED AND ACTUAL BUDGET

Team	Estimated Budget	Actual Budget
Mechanical	$2000	$640.67
Circuitry	$500	$1100.72
Software	$500	$0.00
Total	$3000	$1741.39

III. SYSTEM REQUIREMENTS

The Project aimed to develop an inexpensive multi-purpose space exploration rover system that is capable of image capturing, rock mining, and data collection. Many researchers and engineering teams [29–41] worked on developed new exploration technologies for Moon and Mars applications.

This design effort started by gathering and deriving the requirements from NASA RMC competition rules and regulations as a benchmark. These requirements were followed and frequently checked to meet the competition regulations and goals. As system engineers, the team split the project into functional subsystems and identify their interaction as they are the base of the generated concepts and allowed to create a scoring rubric with respect to meeting the requirements. The main requirements are listed in Table III.

TABLE III. SYSTEM REQUIREMENTS EXTRACTED FROM [23-25]

Requirement Type	Action	Specifications
Performance Requirements	Excavate Regolith	Excavate an adequate depth to reach the ice simulant
	Collect Regolith	Storage to collect the excavated regolith
	Deposit Regolith	Deposit the collected regolith onto a bin located at end of the simulated terrain
Design Requirements	Dimensions	Maximum measurements of 1.5 m in length, and 0.75m in both height and width
	Weight	Maximum weigh of 80 Kg

IV. DUSTYTRON 3.0 PRELIMINARY DESIGNS AND IMPROVEMENTS

The main goals of any systems engineer are continuous improvement and performance enhancement; therefore, the team started by analyzing DustyTRON 2.0 robot and evaluating its performance. DustyTRON 2.0 had some challenges in the area of electrical motors used for wheel and mobility and steering limitations. Hence, DustyTRON 3.0 mechanical team invested significant time to redesign the wheels and steering system to develop various alternatives such as changing the wheel design, acquiring stronger motors with high torque to handle the robot weight, and implementing a new steering mechanism. Additionally, the software team tackled the current code by providing a cleaner and more functional code for both Arduino and controllers, while the circuitry team focused on enhancing the electrical circuit, motor drivers, cable management, wiring harnesses, layout, and power management.

A. Design Development

1) Design 1: This design utilizes an auger to excavate the simulated Martian regolith and a Plexiglas box as storage, as shown in Fig. 3. The frame will be built using T-slotted beams, and PVC pipe, within the following dimensions 1.4 m (length) and 0.75 m (width and height).

The auger system has an auger (0.513 m length, and 0.152 m diameter) and will be enclosed using a PVC pipe (0.152 mm ID). The auger will be attached to a fabricated V-shaped bracket that will be rotated on a pivot point, using two 24-volt heavy-duty servo motors. The home position of the auger is laying inside the collection box. Plexiglas box is sized and positioned to allow the auger to dump and store the regolith. This box will be dumped using 12" linear actuators. Plexiglas was used for the collection box because of its durability and lightweight and has been proven to be able to contain the regolith and protect other components quite efficiently. Electrical boxes will be mounted to the side of the collection box. All four wheels will be 16-inch diameter and 4-inch-wide powered with high torque motors.

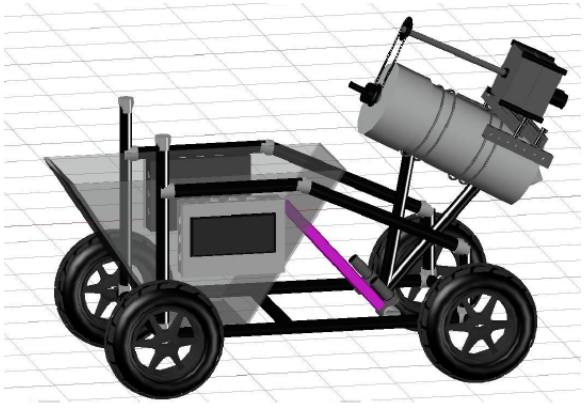


Fig. 3. DustyTRON 3 Mechanical Structure Design 1.

2) Design 2: The structure design 2 close to the DustyTRON 2.0 with few necessary changes as shown in Fig. 4. The team decided to locate the electrical boxes to the sides of the structure while keeping the auger angle fixed and increasing the collection box by modifying the conveyor belt system. The conveyor belt system changed to follow an L-shape, which increases the collection box size while being able to move the regolith from the bottom to the dumping point behind the robot.

3) Design 3: As shown in Fig. 5, the robot design had been developed to include significant modifications such as the dumping system which consists of a single inclined conveyor belt but longer so it can go beyond the rear wheels. Each set of two wheels (front and back wheels) will be attached to a perforated steel tube to create the steering system, this will be attached to the middle frame using two linear bearings and two linear actuators allowing to adjust the height of the robot when needed. In addition, the previous bulky wheels had been eliminated in favor of lighter thinner wheels that would perform the same job.

4) Final design: Various major changes had been conducted to improve the team's design as can be noticed in Fig. 6. The first and most important change is the auger system, which became independent of the middle structure as its tilting angle can be changed using two linear actuators while it can slide down using another linear actuator. Also, wheel-motor attachment has been designed and 3D printed in TAMIU facilities. The team began to build this robot for many reasons; the most important one is the weight, and structural rigidity, and stability which was achieved by using the lightweight T-slotted 80/20 bars. Additionally, the sliding mechanism allowed the auger and wheels to move easily without affecting the frame integrity.

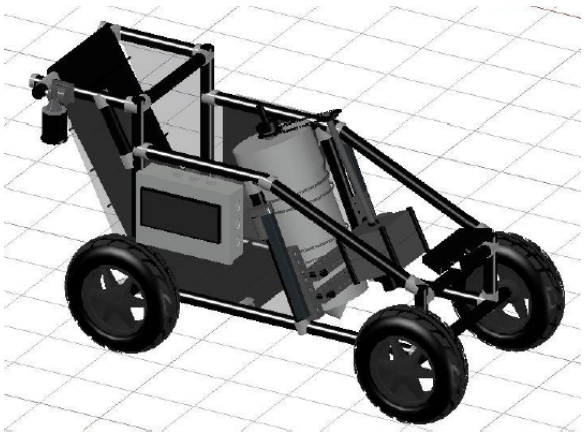


Fig. 4. DustyTRON 3 Mechanical Structure Design 2.

Fig. 5. DustyTRON 3 Mechanical Structure Design 3.

DustyTRON 3.0 requires four (4) independent wheels, ten (10) linear actuators, and one (1) 6-inch inner-diameter excavating auger, which was the foundation of the circuit design. To be able to power all these components with sufficient power distribution and move the 80Kg robot, the team decided to use the following: four 24-Volt motors for the wheels, two 24-Volt motors for the auger, which was confirmed at the testing stage, where power was enough to rotate the auger at the desired speed and using LiPo battery was the best power source for the system.

With the new DustyTRON 3.0 design, a new software configuration and code had been developed to allow simpler

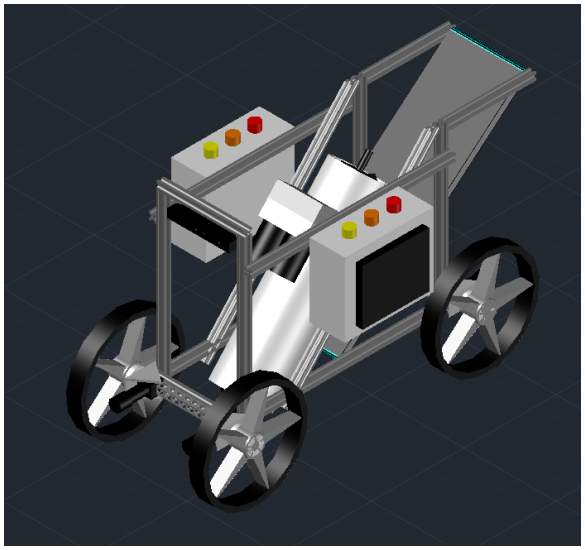


Fig. 6. DustyTRON 3 Mechanical Structure Final Design.

operation and control of the robot. The software team started by reforming the Arduino codes to be easier to read and user-friendly, which allowed easy and precise executable commands utilizing Arduino open-source libraries for the robot individual components, such as the servo library to control Axis 206 Network Camera. While the Pololu motor driver will be using the previously developed library to control all motors and linear actuators.

The changes in the design of DustyTRON 3.0 are now profoundly different, ranging from the wheels to the angle of the auger. The team pushed that boundary of what can be done and showed how much System Engineers can improve on already proven designs.

B. DustyTRON 3.0 Improvements

1) Mechanical team improvements: The main focus for the mechanical team is to fix the steering system of the DustyTRON 2.0 robot, which consists of the wheels, wheel attachment, and motors. While aiming to solve that issue, the team wanted to keep the four-wheel drive (4WD) options as it helps the robot to overcome any obstacle such as a rock or a crater. After extensive research, the team had to find a solution or a method to attach the motor to the middle bar which will act as the rack/structure of the steering system. This bar is a hollowed steel bar that will allow the team to have a strong structure while keeping it lightweight. The original idea required the use of two-wheel casters per side to make a pivotal point and the motor will be mounted in between both wheel casters. The four rods attached to the wheel caster shown in Fig. 7 will be used to securely attach the motor in place.

After detailed analysis, this idea had been developed to include a square perforated steel tube, this new design will use two-wheel casters only and the tube will be used to mount the linear actuator brackets. Fig. 8 shows the steering system with the perforated steel tube and linear actuators. With one caster at each end, the motor can be attached directly to the inner vertical wall of the caster, which means less mechanical interference and easier rotational motion for steering. The shaft would be the only thing coming out of the caster which connects both wheels and will be connected to the robot structure.

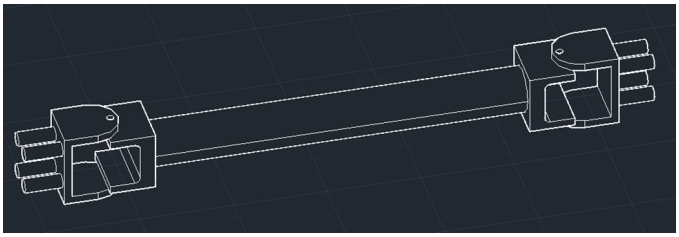


Fig. 7. DustyTRON 3 Proposed Steering System.

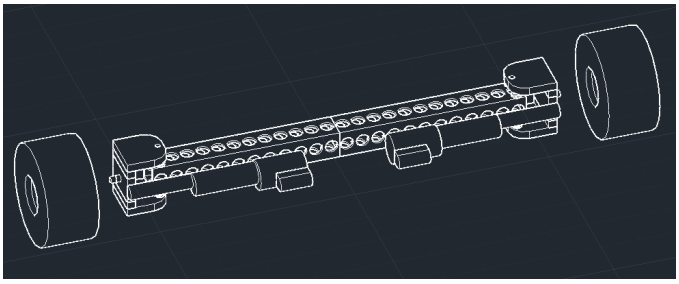


Fig. 8. DustyTRON 3 Final Steering System.

After finalizing the motor attachment, the team found a solution to attach the motor shaft to the wheel by building a two-parts wheel hub. This new hub consisted of an Alumni 8mm screw hub which will be attached directly to the motor shaft. While the second part will be a special part that will match with the grooves of the wheel and both parts will be attached together using bolts and locknuts. Unfortunately, the team did not have access to the Computer Numerical Control (CNC) machines to manufacture the complete wheel hubs; therefore, the team decided to utilize the 3D printer to create the hub-wheel attachments. These hubs went through various phases until a perfect fit was found, Fig. 9 shows the different phases/designs of the wheel hubs.

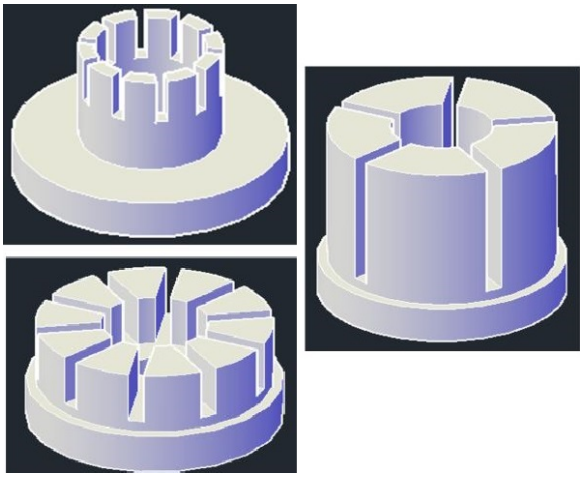


Fig. 9. DustyTRON 3 Wheel Hubs.

DustyTRON 3.0 steering system will be very efficient and

have easier maneuverability compared to the previous robot steering, which will reduce the time required for steering in the competition runs. In addition, springs and wheel bearing had been added to the caster in order to add more flexibility and enhance the rotation movement of the caster as shown in Fig. 10.

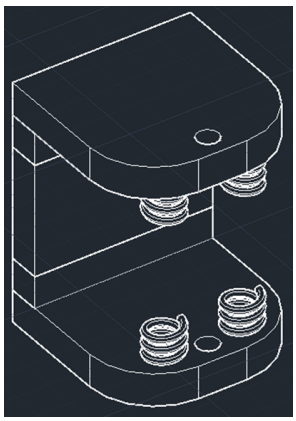


Fig. 10. DustyTRON 3 Caster Design.

The selected wheels for the final design, were $14inches$ in diameter and $1.75inches$ in width. This will help the team to have a smaller footprint while steering and will allow the team to decrease the robot turning angle.

Another significant improvement is that DustyTRON 3.0 will contain a suspension system that will control the height of the robot using linear actuators, which will change the clearance between the auger tip and the ground from $4inches$ to $10inches$. This height difference would allow the robot to go over small to medium-sized obstacles. Several designs were taken into consideration before selecting the final design. One more improvement to the mechanical structure is providing more rigidity to the auger structure; this was achieved by adding two t-slotted bars to join the front and back ends of the robot. Two linear actuators had been added in order to change the auger tilting angle and reduce the vibration transmission to the robot structure.

2) Circuitry team improvements: The circuitry team focused on solving the electrical components layout and wiring issues. The team decided to purchase higher quality electrical boxes, design a better component layout within the component boxes, better component box placement onto the robot and decided to install fans within the motor control box. Having that in mind, the team started working on designing the layout within the component boxes to organize the cables to harness the wires in a way to reduce the cable length. This reduces the risk of having an excess of loose cables and reduces the electrical noise that might affect the microcontroller and motor driver's performance. These component boxes will house and protect all the electrical components that are required for the robot operation, which resulted in two-component boxes that will be explained below.

Electrical Box 1: Main Brain Box The first electrical box is labeled Main Brainbox since is considered the main computer of the robot. It consists of a 14.8V Lithium Polymer RC Battery, power on/off switch, power analyzer, fuse, 1 E-Stop button, voltage step down, NVidia TK1 will be connected to LAN line, USB HUB, and display with HDMI cable. The NVidia TK1 will be powered through a step-down voltage regulator as it requires 12 volts. The USB HUB will be plugged in the NVidia USB 3.0 port to power the Arduino, a rearview camera, keyboard and mouse, and a data terminal for the Xbox Kinect camera. Fig. 11 shows the battery, power switch, power analyzer, E-Stop, fuse, step down, and NVidia connection in this box.

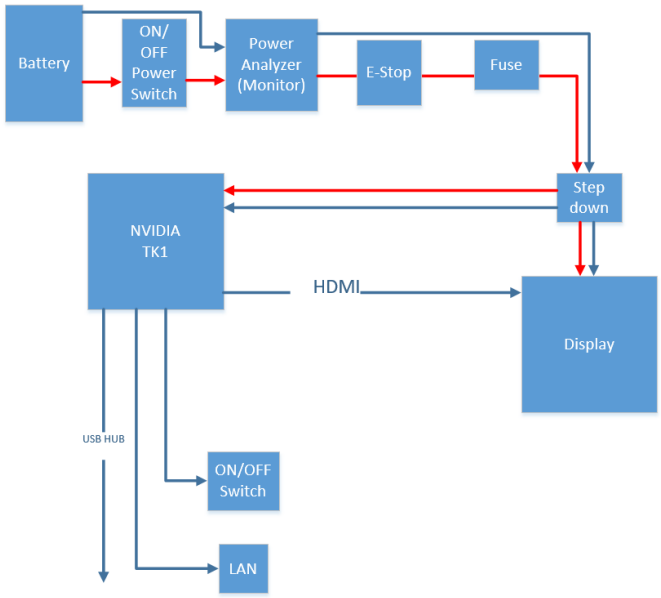


Fig. 11. DustyTRON 3 Robot - Main Brain Box: Component Circuit Diagram.

Electrical Box 2: Motor Control Box Electrical box number 2, labeled as Motor Control Box, is where all the other electrical components are connected to all hardware components that are vital for the robot's physical movement. This box contains 1 power switch, 1 power analyzer, 1 emergency stop button, 1 battery, 1 eight fused-output power distributor board, 3 fans, USB HUB, 1 operation flashing light, 1 Arduino, and 6 motor drivers. The 22.2 V battery is connected in series to the emergency stop button, which is then connected to the power analyzer that is connected to the power switch. The emergency stop button and the fuses in the power distributor work as a method of safety to protect the circuit from any malfunctions. The operation flashing light is connected directly to the power distributor and it will be used as an indication of robot reediness. The power distributor contains an on and off switch and eight fused power outlets. The terminals to the motor drives are connected to six of the power distributor outlets. One of these fused power outlets will be connected to a step-down voltage regulator that had been adjusted to output 12 volts only to power the cooling fans, and a second step-down voltage regulator will be also adjusted to 12-volt output to power the Kinect.

The purpose of the fans and the heat sinks attached to the motor drivers is to extract the heat created within the motor control component box. Two of the six motor drivers will be connected

to each motor on the wheels, one motor driver for the front wheels and one motor driver for the back wheels. These motor drivers have dual channel connections. The other motor drivers will be used for the steering and suspension actuators, and the slider actuators and conveyor belt. These Pololu motor drivers are dual channels that are capable of delivering up to 12A each. Therefore, by combining both channels into a single, that will have provided the motor attached with current up to 24A. The auger and horizontal auger actuators are connected to be in single-channel mode. Fig. 12 illustrates the components and the connection in the motor control box.

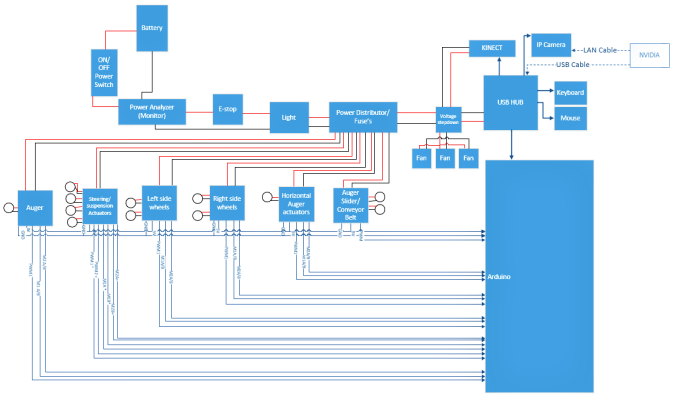


Fig. 12. DustyTRON 3 Robot - Motor Control Box: Component Circuit Diagram.

These electrical boxes are made of thick plastic in comparison to the previous robot's boxes that were made out of plexiglass. These boxes provided the team with a better structure to mount the E-stop and power switches without influencing the integrity and weather resistance of the boxes. These boxes will be mounted in the available space between the wheels and on the sides of the robot, with removable mounting brackets. This location will provide better weight distribution and having the electrical boxed in an elevated position, which will keep the components safe far from highly active moving components and allow easier access for maintenance, troubleshooting, and even parts replacement when needed.

Another improvement was the addition of a fused power distributor board that can handle the supplied voltage and current; since this will regulate the voltage into the motor drives, operation flashing light, E-stop, power switch, and voltage step down. Lastly, the circuitry and software team decided to install a 10-inch display monitor, that can be attached to the main electrical box that houses the NVidia, which will make troubleshooting easier.

3) Software team improvements: The software team focused their effort on creating an optimized code that will be used to control all the linear actuators and motors within the robot. In addition, the software team worked on enhancing the code for the manual control mode and provided a cleaner, simplified, and user-friendly code that will be used to interface the Xbox controller to the NVidia microprocessor. Optical sensors within Xbox Kinect and a servo camera were added to the robot for autonomous operation and obstacle detection and avoidance. The Kinect camera was placed at the font to frontal environment scan, while the servo camera is was mounted in the back of the robot to monitor and regulate the deposit mechanism. DustyTRON 3.0 used the Jetson TK1 Graphics Processing Unit (GPU) to facilitate autonomous operation by implementing wireless communication and computer vision.

V. Concept Operations

DustyTRON 3.0 will have a better steering system, new wheels, and a new overall structure design. The weight will decrease due to battery change and the frame will still be constructed by $80/20$ T-Slotted bars because of the lightweight and easy manipulation. To excavate the simulated Martian terrain, DustyTRON 3.0 will use a double helix auger that is powered with dual motors with a gearbox of $47:1$ ratio and two linear actuators to move the auger into the ground. DustyTRON 3.0 has $14inches X 1.75inch$ wheels in order to have better steering and powered with 24 V high torque motors with a $295:1$ great box ratio. The suspension mounted on the structure of the robot is meant to lift the robot, in order to go over the larger rocks on top of the first layer of BP-1 and to lower the robot once the digging process starts. DustyTRON 3.0 was redesigned to operate autonomously with the use of two cameras, a microprocessor, and a graphical processing unit. The designed autonomous mode utilizes a Microsoft Xbox 360 Kinect camera with an IR sensor that provides video data to the CPU, and the NVidia TK1, for object detection. The second rear servo IP camera provides the team video data that regards to regolith deposit.

In case of autonomous mode failure, DustyTRON 3.0 can be controlled over WiFi by two Xbox 360 controllers from a max distance of about 50 feet. This manual control of the robot is established by sending simple 8-bit commands that resemble keyboard strokes, to an Arduino Mega 2560 unit that is directly connected to an NVidia TK1. Whether the robot is in manual or autonomous mode, NVidia TK1 will pass the commands to the Arduino that directly controls all motor drivers and operate the wheel motors and actuators.

VI. System Hierarchy

The relationship within the main components of each subteam can be illustrated using a system hierarchy diagram. For the mechanical team, the moving, excavating, and deposit system relationship is shown in Fig. 13. Circuitry and software sub-teams share several common components such as controller, NVIDIA GPU, Arduino, and motors. Both subteams will be working to assure both manual and autonomous operation of the robot. Manual control will be based on the pictures and live feeds provided by Xbox Kinect, the human operator will use an Xbox 360 controller to send the command through the controller computer to the robot's NVIDIA TK1 processor, which are interconnected through wireless WiFi and use SSH (Secure shell). This SSH was selected as it is an encrypted network protocol, and will prevent any unauthorized access to the robot's TK1.

In the autonomous mode, the Xbox One Kinect feed will be processed directly within the TK1 for object detection purposes, while the rear camera feed will be used for the regolith collection and deposit process. In both scenarios; manual or autonomous mode, TK1 will command the Arduino Mega, which is directly controlling the motor drivers and

mechanical components. Fig. 14 shows the circuit hierarchy and Fig. 15 shows the software system hierarchy.

Additionally, the software team implemented a VNC communication to reduce the used bandwidth by compressing the video feed before broadcasting to the main control station. Within the robot system, serial communication and powering the Arduino was done using the USB port, which simplified the Arduino power circuit.

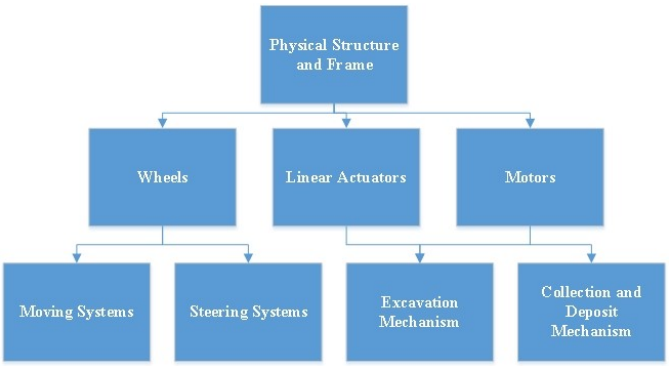


Fig. 13. Mechanical System Hierarchy for DustyTRON 3.0.

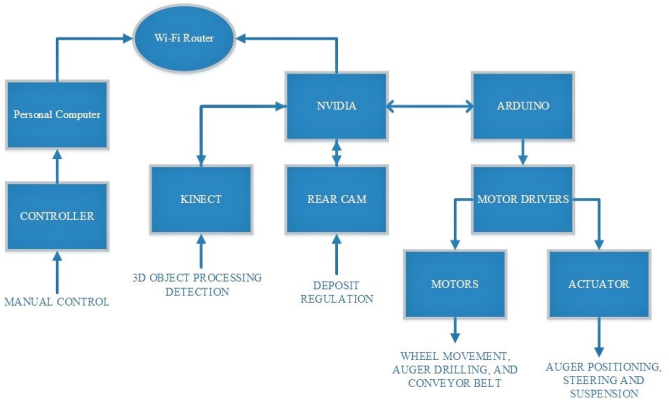


Fig. 14. Circuitry System Hierarchy for DustyTRON 3.0.

VII. Robot Interface

One of the main objectives of the DustyTRON 3.0 engineering design process was to create a reliable and maintainable interface between our subsystems. Fig. 16 shows a level diagram for components' interface for the mechanical, electrical, and software systems.

The robot interface was built using Ubuntu 14.04 as the NVidia TK1 operating system (OS) with different software such as Arduino Software IDE to communicate the Arduino Mega, Microsoft XNA to program the Wired Xbox 360 controllers that will be used for the robot manual control mode by developing a Visual Basic (VB) Code, while PuTTY was used for serial port communication between the NVidia TK1 and the Arduino to emulate Arduino's serial console to receive input data, and TurboVNC were used to establish two-way secure remote communication [42].

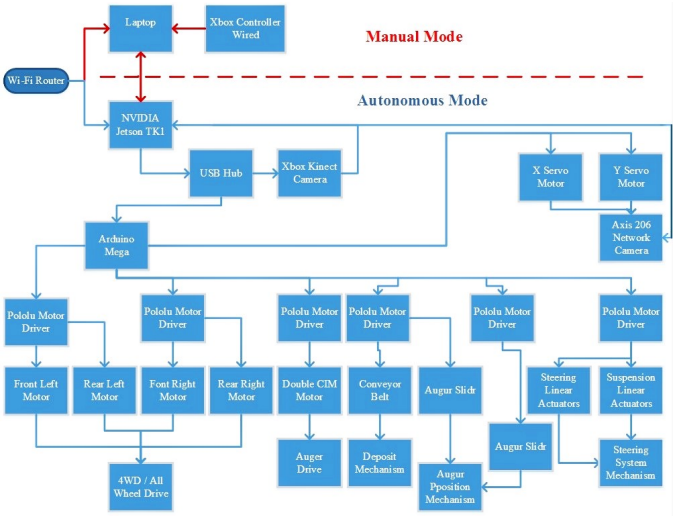


Fig. 15. Software System Hierarchy for DustyTRON 3.0.

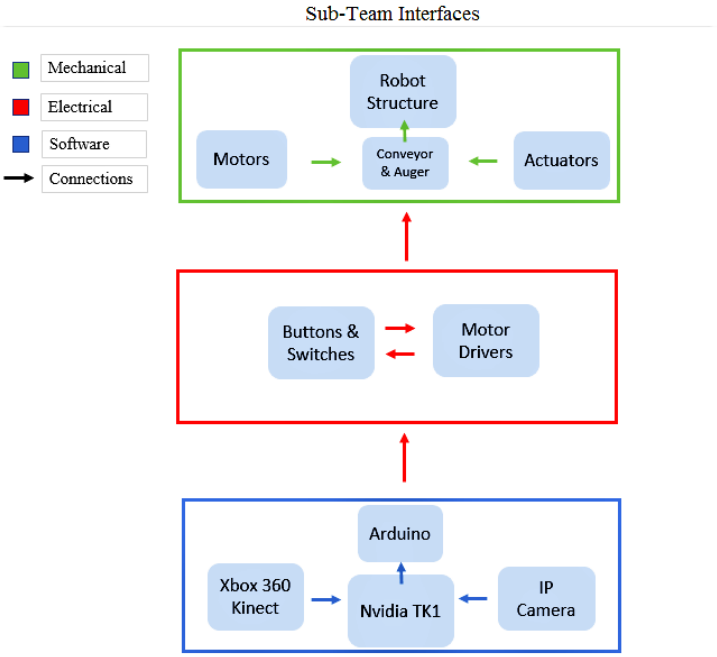


Fig. 16. Systems Interface for DustyTRON 3.0.

VIII. Risk Management

As a Systems Engineer, prediction and planning for the future is an essential task in any project. This step can be achieved by analyzing the system of interest for any possible failures and developing a ranking system for failure consequences on the overall performance of the robot and likelihood of happening, which will allow the team to adapt and prepare with a solution if issues arise. Each sub-team developed their risk matrices and they are as follow:

A. Mechanical Team

The Mechanical Team analyzed the mechanical structure and components and found the following risks that might occur

during or before the competition:

- **Failure to move**: if the steering system cannot function, or fails to traverse to excavate. **Major Consequences and Unlikelihood of Occurrence**
- **Failure to excavate**: if the excavation (Auger) system is not working as expected due to a failed motor or chain or obstacles exist within the system. **Catastrophic Consequences and Rare Likelihood of occurrence**
- **Failure to collect regolith**: in the event where the collecting mechanism fails to hold regolith or the excavation mechanism is not providing enough regolith. **Major Consequences and Moderate Likelihood of Occurrence**

B. Electrical Circuitry Team

The risks that might occur during or before the competition were found to be:

- **Failure of Circuitry**: if the circuitry/cables fail and burn due to an unexpected overheat, which might lead to the entire component box failing. **Major Consequences and Rare Likelihood of Occurrence**
- **Failure of Battery**: if batteries failed to hold an electrical charge or not able to provide the required electrical power. **Moderate Consequences and Moderate Likelihood of Occurrence**
- **Failure of Motor Drivers**: if motor drive overheats or stops responding to the Pulse Width Modulation (PWM) signal. **Major Consequences and Unlikelihood of Occurrence**
- **Failure of Motors**: if the motor malfunctions and not able to rotate the attached mechanical component. **Catastrophic Consequences and Unlikelihood of Occurrence**

In the building phase, the circuit team was actively testing and verifying the proper operation of every single component to prevent any future issues. They simplified the circuit design to allow fast and easy components' diagnostic and replacement.

C. Software Team

For the software architecture design, some failures can be due to connection with a mechanical-related failure. The major risks that had been considered are as follows:

- **Failure of feedback**: if the connection fails to send feedback on possible problems or updates, or if the Kinect camera or IP camera loses signal, the team will be prevented from viewing the terrain. **Minor Consequences and Unlikelihood of Occurrence**
- **Failure of NVIDIA TK1 power regulator**: if the power circuit fails to provide the TK1 with the required 11.6-12.6 Volt, then the GPU will fail and go into limp mode. **Moderate Consequences and Unlikelihood of Occurrence**
- **Failure of VNC connection**: If the remote access connection fails to be established, then the robot won't be controlled manually or it might not be able to receive the autonomous start signal. **Catastrophic Consequences and Moderate Unlikelihood of Occurrence**
- **Failure of programming OpenCV**: if the vision and image analysis system fails to start then the robot won't run autonomously. **Major Consequences and Rare Likelihood of Occurrence**
- **Failure to send a command to Arduino**: if serial communication fails between the TK1 and Arduino, then the robot will fail to do the required mission as no movement will be executed. **Catastrophic Consequences and Rare Likelihood of Occurrence**

IX. Trade-off Assessments

Using Quality Functional Deployment (QFD) method, the team was able to find their design's strength and worked on enhancing them using a trade-off assessment for every subteam.

A. Mechanical Team Trade-off Assessment

The robot's mechanical structure strength was found to be the independent controlled suspension, which will allow the robot to adjust its height to go over obstacles and rough terrain while adding extra components and weight to the robot.

B. Circuitry Team Trade-off Assessment

The previous design of the robot was based on utilizing six (6) 12V 7Ah sealed batteries, which were 4.5lb each, the team decided to switch to two LiPo batteries (14.8V and 24V) which are lighter (2.6 lbs in total) but extremely powerful and careful circuit design is required to prevent any damage to electrical components such as TK1 or motor drivers. The circuit was improved by changing wires to thicker gauge (12 AWG), adding heavy-duty power distribution with fused ports, including fans to extract the heat within the electrical boxes, and batteries were protected using Fireproof Safe Bag. In addition, the VNH5019 Pololu motor drives were used although they require soldering and extra configuration for mono or dual channel. They provided superior performance and accuracy to a single motor or double motors control.

C. Software Team Trade-off Assessment

For manual control implementation, two wired Xbox 360 controllers were used to eliminate the wireless connection lag and it will allow the simplification and splitting of robot controlling tasks by having one person control the excavation system and another operate the robot mobility. The team chose to implement autonomy by using an Xbox Kinect Camera. If autonomy fails, the team will change to manual control to regain control of the robot. The Xbox Kinect camera was used instead of the PS3 Move camera. Although the PS3 Move camera used less power consumption and had a better resolution, the Xbox Kinect had integrated sensors that will allow the 3D mapping feature needed for the autonomous mode [43]. These integrated sensors balanced the power consumption

consequence since having to get extra individual components would result in a similar outcome. Arduino Mega 2560 [44] was selected to control and command the motor drive utilizing its superior and stable PWM compared to the Jetson TK1. Axis 206 Camera was used for excavation and collection operation monitoring as it has two servo motors to control its aim in X and Y directions.

X. Requirements Verification

To assure that the robot was designed and built to meet NASA regulations, the following requirements were checked frequently:

A. Functional Requirements

- Robot must traverse the simulated Martian terrain and excavate the needed regolith from the mining area.
- Tele- or autonomous operation of the robot.
- Sufficient size collection system to stored Regolith until the deposition.
- Obstacle avoidance in the arena.
- The robot's suspension shall be able to lift the rear or front end as desired.
- Robot code must be simple and easy to execute.

B. Performance Requirements

- The robot shall be able to start the mission from any assigned location or orientation.
- Collect and deposit $10kg$ of BP-1 within the allowed 10-minute mission.
- Excavate BP-1 from the designated area only.
- Dust prevention and electrical components protection.
- Limited bandwidth and power consumption.

C. Physical Requirements

- Maximum weigh of $80kg$.
- Self-sustained power with consumption monitoring and recording system.
- Initial dimensions of $1.5m$ Length, $0.75m$ width, and $0.75m$ height.

D. Safety Requirements

- An emergency stop red button with a diameter of $40mm$ in an easy and safe accessible position.
- All wire harnesses are securely attached and protected.
- Easy and secure connection to the robot control systems.

To assure that the robot is meeting all NASA RMC's requirements, the team verified their design in the testing phase by inspecting the Commercial off-the-shelf (COTS) parts used to perform as intended. Some of the inspected items are shown below:

- The extruded T-slot bars' integrity was inspected under loading and vibration conditions.
- Operate the twin spiral auger in a simulated sand field to assist its performance and measure the collected sand weight.
- Linear actuators and motors were tested before and after fitting it to the frame to make sure of their ability to move the robot.
- The conveyor belt system was tested to check its operation and ability to move the regolith from the collection box to the dumping location.
- Batteries were charged and monitored to guarantee they can last for the 15-20 minutes mission.
- Emergency-stop buttons were tested where the power to the whole robot was shut down safely.
- Axis 206 and Xbox Kinect camera functionally were tested.

XI. Reliability

To ensure the robot's maximum reliability of the robot, a few actions had been taken:

- Hardware team strengthening the structure and reduced the weight, improved the wheels, steering, and suspension systems to overcome the harsh terrain.
- Circuitry team arranged the electrical components to reduce and manage the cables and connections effectively, and by using LiPo batteries their contributed to reducing the robot's total weight.
- Software team utilized the serial communication and VNC connection secure the interface between NVIDIA, Arduino, and main control computer. Additionally, codes were updated and improved and Fig. 17 shows the updated pseudocode for the Arduino.

XII. Competition Results Summary

The DustyTRON robot shown in Fig. 18 was delivered to Kennedy Space Center in Florida to participate and compete against 53 robots from all over United State. The robot passed all the inspection procedures and after the competition runs, it placed the 29th and was awarded the "Innovative Design" for its unique steering and suspension systems. This experience was exceptional, which allowed the team members to show their engineering skills and participate in the race of space exploration.

XIII. Conclusion

DustyTRON team members represented in Fig. 19 showed their skills an interesting real-life challenge. Placing the 29th out of 54 invited universities and getting the "Innovative Design" Award, were a great conclusion for the third-year team performance. The team designed and build a unique robot that represents the mechanical, electrical, and software constraints, by being resourceful and implementing systems engineering principles to solve world-level problems.

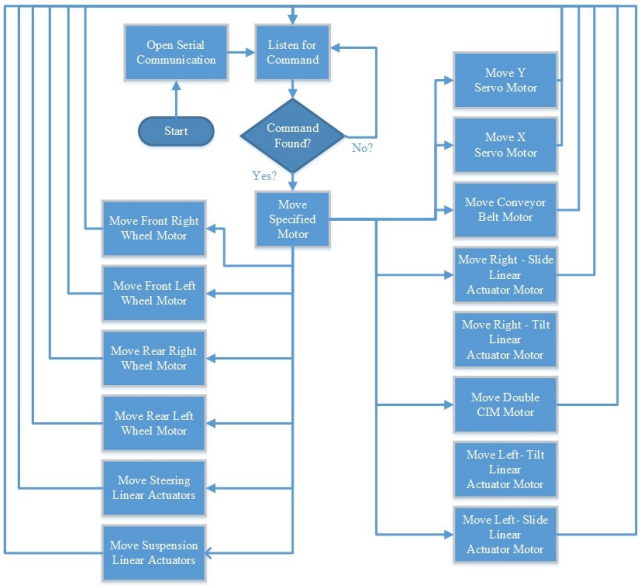


Fig. 17. DustyTRON 3 Robot - Arduino Code Flow Chart.

Fig. 18. DustyTRON 3 Robot - Final RMC 2017.

The team strives to improve on team management, time management, and leadership skills. The team will continue seeking new mechanical, circuitry, and software-related concepts to be implemented on future robot designs. Throughout the project, the team has been able to implement engineering skills acquired as Systems Engineering students but also learned how to work as a cohesive team while adding new skills.

For future competitions, the team will practice the continuous improvement principles to learn from their mistakes and develop a better robot:

- Enhance the autonomous operation by utilizing new computer vision algorithms.
- Improve the excavations system to include auger and conveyor system.
- Enhance the adaptive suspension and steering systems to be more compatible with harsher terrain.

The team was inspired to share this experience with local students and the community, hosting STEM days at local schools promoting the interest in the robotics field and NASA's programs and projects. Additionally, the team took advantage of all the venues to support and mentor local FIRST Lego/Tech/Robotics teams and host their local competitions.

Fig. 19. DustyTRON 3 Team Members.

ACKNOWLEDGMENT

Thanks to NASA for providing such a great opportunity to participate in such a competition. This project would not be successful without the support of Texas A&M International University (TAMIU) and our sponsors from Laredo, TX.

REFERENCES

[1] Voosen, Paul. "Mars rover steps up hunt for molecular signs of life." (2017). Science 03 Feb 2017, Vol. 355, Issue 6324, pp. 444-445, DOI: 10.1126/science.355.6324.444.

[2] Koris, Daniel R., and Jason Isaacs. (2017) "A Formal Approach to Extended State Machines for Multi-Objective Robots Operating in Dynamic Environments." Proceedings of the 2017 Midstates Conference on Undergraduate Research in Computer Science and Mathematics.

[3] Tashtoush, T., Velazquez, A., Aranguren, A., Cavazos, C., Reyes, D., Hernandez, E., Bueno, E., Otero, E., Zamudio, G., Casarez, H., Rullan, J., Rodriguez, J., Villarreal, J. C., Gutierrez, M., Rodriguez, P., Torres, R., Martinez, R., and Partida, S., "Developing a Mining Robot for Mars Exploitation: NASA Robotic Mining Competition (RMC)", International Journal of Advanced Computer Science and Applications(IJACSA), 11(12), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111205.

[4] Tashtoush, T., Hernandez, R., Yanez, R., Gonzalez, J., Moreno, H., and Escobar, V. (2020). "Reverse-Twister Swarm Search Algorithm Design: NASA Swarmathon Competition", International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 7(1), pp.13-20.

[5] Hernandez, R., Yanez, R., Gonzalez, J., Moreno, H., Escobar, V., and Tashtoush. T., (2016) "Design of a Swarm Search Algorithm: DustySWARM Reverse-Twister Code for NASA Swarmathon." Texas A&M International University, School of Engineering.

[6] Tashtoush, T., Gutierrez, O., Herrera, E., Medina, J., Peña, A., Varela, E., and Hernandez, R. (2020). "Design of a Swarm Search Algorithm: DustySWARM Spiral Epicycloidal Wave (SEW) Code for NASA Swarmathon", International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 7(1), pp.28-36.

[7] Gutierrez, O., Herrera, E., Medina, J., Peña, A., Varela, E., Hernandez, R., and Tashtoush. T. (2017) "Design of a Swarm Search Algorithm: DustySWARM Spiral Epicycloidal Wave (SEW) Code for NASA Swarmathon". Texas A&M International University, School of Engineering.

[8] Tashtoush, T., Ruiz, C., Estevis, T., Herrera, E., Bernal, R., Martinez, R., and Reyna, L. (2020). "Square Spiral Search (SSS) Algorithm for Cooperative Robots: Mars Exploration", International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 7(1), pp.21-27.

[9] Ruiz, C., Estevis, T., Herrera, E., Bernal, R., Martinez, R., Reyna, L., and Tashtoush, T., (2018) "Design of a Swarm Search Algorithm: Square Spiral Search (SSS) Algorithm for NASA Swarmathon". Texas A&M International University, School of Engineering.

[10] Tashtoush, T., Ahmed, J., Arce, V., Dominguez, H., Estrada, K., Montes, W., Paredez, A., Salce, P., Serna, A., and Zarazua, M., "Developing a Radiating L-shaped Search Algorithm for NASA Swarm Robots", International Journal of Advanced Computer Science and Applications (IJACSA), Volume (11), Issue (8), August 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110802.

[11] Ahmed, J., Arce, V., Dominguez, H., Estrada, K., Montes, W., Paredez, A., Salce, P., Serna, A., Zarazua, M., and Tashtoush, T. "Developing a Radiating L-shaped Search Algorithm for NASA Swarm Robots", Texas A&M International University, School of Engineering.

[12] Secor, P. (2016). "NASA Swarmathon".

[13] Braccio, M. (2019). Design of a Robot for the 2019 NASA Robotic Mining Competition. In Proceedings of the Wisconsin Space Conference (Vol. 1, No. 1).

[14] Neubert, J. J. (2016). Using NASA's Robotic Mining Competition to Give Students a Quality Systems Engineering Experience. In ASEE's 123rd Annual Conference & Exposition (pp. 4-11).

[15] Guerra,L., Murphy, G., and May. L., (2013). Applying Engineering to the Lunabotics Mining Competition Capstone Design Challenge. Proceeding of the ASEE Annual Conference and Exposition, June 2013.

[16] Stecklein, J. (2017, July). NASA's Robotic Mining Competition Provides Undergraduates Full Life Cycle Systems Engineering Experience. In INCOSE International Symposium (Vol. 27, No. 1, pp. 1456-1473).

[17] Berrios, D. C., Galazka, J., Grigorev, K., Gebre, S., & Costes, S. V. (2020). NASA Genelab: Interfaces for the Exploration of Space OMICS data. Nucleic Acids Research, 2020 Oct 20. https://doi.org/10.1093/nar/gkaa887.

[18] Balint, T. S., Kolawa, E. A., Cutts, J. A., & Peterson, C. E. (2008). Extreme environment technologies for NASA's robotic planetary exploration. Acta Astronautica, 63(1–4), 285– 298. https://doi-org.tamiu.idm.oclc.org/10.1016/j.actaastro.2007.12.009.

[19] Bogue, R. (2012). Mars curiosity: sensors on the red planet. Sensor Review, 32(3), 187–193. https://doi-org.tamiu.idm.oclc.org/10.1108/02602281211233151.

[20] Cohen, B. A., Chavers, D. G., & Ballard, B. W. (2012). NASA'S Robotic Lunar Lander Development Project. Acta Astronautica, 79, 221–240. https://doi.org.tamiu.idm.oclc.org/10.1016/j.actaastro.2012.03.025.

[21] Cole, T. J., Bassler, J., Cooper, S., Stephens, V., Ponnusamy, D., Briere, M.,& Betenbaugh, T. (2012). The challenges of designing a lightweight spacecraft structure for landing on the lunar surface. Acta Astronautica, 71, 83–91. https://doi-org.tamiu.idm.oclc.org/10.1016/j.actaastro.2011.08.003.

[22] Chaput, A., 2016, 'System Engineering Education for All Engineers - A Capstone Design Approach'. ASEE 123rd Annual Conference & Exposition, New Orleans, June 26-29, 2016.

[23] Mahmood, M. 2016, 'Oakton Community College 2016 NASA Robotic Mining Competition Systems Engineering Paper', paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[24] The University of Alabama in collaboration with Shelton State Community College, 2016, 'Journey to Mars; 2016 Systems Engineering Paper', paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[25] Charlotte 49er Miner Robotics, The University of North Carolina at Charlotte, 2016, '2016 Systems Engineering Paper', paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[26] Illinois Robotics in Space (IRIS), the University of Illinois at Urbana-Champaign, 2016, 'Design and Development of the IRIS-6 Robotic Mining System', paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[27] John Brown University Eaglenaut Robotics, John Brown University, 2015, 'Robotic Regolith Excavation System', paper presented to the 2015 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 18-22 May.

[28] Chicago EDT Robotics, the University of Illinois at Chicago, 'Systems Engineering Report 2016, the University of Illinois at Chicago, AMES-3 – Surus', paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[29] Iowa State University Cyclone Space Mining, '2015-2016 Systems Engineering Paper', paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[30] Dieter, G. E., & Schmidt, L. C. (2013). Engineering design. Boston: McGraw-Hill Higher Education.

[31] "Rules and Rubrics", Nasa.gov, 2017. [Online]. Available: http://www.nasa.gov/offices/education/centers/kennedy/technology/nasarmc/RulesRubricsResources.

[32] Kapurch, S. J. (Ed.). (2010). NASA systems engineering handbook.National Aeronautics and Space Administration. Diane Publishing.

[33] Bellestri, S., Boil, T., Carswell III, M., et al. (2013). Alabama Lunabotic 2013 Systems Engineering Paper (Undergraduate Thesis) Retrieved from NASA.

[34] Alfaro, D., Aranguren, A., Duarte, T., De La Cruz H., Perez, G., Torres, A. Delgado, J., Melero, D., Vazquez, J. A., Charlton, B., Jose Guajardo, J., Flores, G., Garza, E., & Tashtoush, T. (2015). Systems Engineering Paper (Undergraduate Thesis) Retrieved from Texas A&M International University School of Engineering.

[35] Ay, N., Bertschinger, N., Der, R., Güttler, F., & Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. The European Physical Journal B, 63(3), 329-339.

[36] Shotts Jr, W. E. (2012) "The Linux command line: a complete introduction" No Starch Press.

[37] Carson, E. M., Rivadeneira, J., Woodward, N. K., & Peterson, P. W. (2016). "NASA Robotic Mining Competition 2015-2016".

[38] Mueller, R. P. (2012) "Lunabotics Mining Competition: Inspiration through Accomplishment" Thirteenth ASCE Aerospace Division Conference on Engineering, Science, Construction, and Operations in Challenging Environments, and the 5th NASA/ASCE Workshop On Granular Materials in Space Exploration.

[39] Williams, W. B., & Schaus, E. J. (2015). Design and Implementation of a Rocker-Bogie Suspension for a Mining Robot. In ASEE Southeast Section Conference.

[40] Liu, Y., Jeremy B., Zachary C., Jennifer B., John A.s, Madelyn D., David S., Cindy L. B., John B., and Christopher A.. "Mechanical design, prototyping, and validation of a Martian robot mining system." SAE International Journal of Passenger Cars-Mechanical Systems 10, no. 2017-01-1305 (2017): 177-182.

[41] Mueller, R., & Van Susante, P. (2011, September). A review of lunar regolith excavation robotic device prototypes. In AIAA SPACE 2011 Conference & Exposition (p. 7234).

[42] A Brief Introduction to TurboVNC. (2016, February 20). http://www.turbovnc.org/About/Introduction.

[43] Mojtahedzadeh, R. (2011). Robot obstacle avoidance using the Kinect. Master of Science Thesis Stockholm, Sweden. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-130746.

[44] Arduino MEGA 2560 & Genuino MEGA 2560. (2017). https://www.arduino.cc/en/Main/arduinoBoardMega2560.

Evaluating the Accuracy of Models for Predicting the Speech Acceptability for Children with Cochlear Implants

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

***Abstract*—This study developed a model for predicting healthy hearing people's speech acceptability for children with cochlear implants using multiple regression analysis, support vector regression, and random forest and evaluated the prediction performance of the model by comparing mean absolute errors and root mean squared errors. This study targeted 91 hearing-impaired children between four and eight years old who had worn cochlear implants at least one year and less than five years. Speech data of children wearing cochlear implants (CI) were collected through two tasks: speaking and reading. The outcome variable, healthy hearing people's speech acceptability for children wearing CI was evaluated by 80 college students (freshman and sophomore) who did not have prior knowledge of children with a cochlear implant. The results of this study showed that the random forest algorithm (mean absolute errors=0.81and root mean squared error=0.108) was the best model for predicting the speech acceptability of children wearing CI. The results of this study imply that the predictive performance of random forest will be the best among ensemble models when developing a machine learning model using speech data of children wearing CI.**

***Keywords*—Cochlear implants; speech acceptability; support vector regression; random forest; mean absolute errors**

I. Introduction

Since the National Health Service of South Korea began to cover cochlear implants in 2005, cochlear implants have become more common for the hearing impaired in South Korea. The ear consists of the external ear, the middle ear, and the internal ear, and the cochlear implantation refers to an operation of implanting an artificial cochlea device in the ear of the patient who cannot hear voice due to the damage of the cochlea to help the patient hear speech [1]. Cochlear implants provide useful hearing for children with severe hearing difficulties or deaf children who cannot hear speech even with hearing aids [2, 3]. Many children with hearing impairments have benefited greatly from cochlear implants (CI) for enhancing their hearing ability and developing language ability [4]. Particularly, the ultimate goal to obtain through cochlear implants is to improve communication skills through vocal language [5]. Consequently, many studies [6, 7] have shown interest in the ability of children to produce spoken language (speech) after cochlear implants, and they have sought ways for children with cochlear implants to produce better speech than before operation based on the improved hearing ability.

Speech intelligibility and speech acceptability have been used widely in the speech-language pathology field to compare speech characteristics and severity for diverse communicative disorders such as articulation and phonological disorders, dysarthria, and apraxia of speech [8, 9]. Among them, speech acceptability refers to how well the content that the speaker is trying to convey is delivered to the listener (how well the listener understands it", and it is mainly used as an index reflecting the success of expressive speech [8]. Since the listener listens to the speaker, various variables (e.g., vocal intensity, pitch, and speech rate) comprehensively influencing the listening. It is necessary to have an index for evaluating the overall speech production ability of the speaker from the listener's point of view [5]. Speech acceptability is used as an index to evaluate the overall speech production ability. Speech acceptability measures how naturally the speaker's intention is understood by the listener, and is a representative index showing the expressive ability of the hearing impaired.

Nevertheless, previous studies measured speech acceptability to mainly evaluate the speech characteristics of patients with dysarthria or the hearing-impaired due to neurological damage such as stroke and to identify the speech characteristics of cleft palate patients [8, 9, 10]. These studies compared the results with the speech acceptability of the healthy control group based on traditional statistical analyses such as t-test and ANOVA [8, 9, 10]. Only a few studies examined the predictors of speech acceptability using machine learning.

The general public has become more interested and gained a better understanding in machine learning in various fields (e.g., finance, medicine, and engineering) [11, 12]. The South Korean government also pays a lot more policy interest in the artificial intelligence (AI) and health care industries. AI refers to a technology for analyzing data and identifying better answers through the visualization of big data, machine learning, and deep learning of big data. Among them, machine learning indicates a prediction technique to predict changes by reading numerous data and discovering hidden algorithms. In the healthcare industry, there have been many cases of applying and utilizing AI technologies including machine learning [13], such as cancer diagnosis and treatment recommendations using AI-based IBM Watson, diagnostic medicine using machine learning analysis techniques, and new drug development systems. Studies have continuously reported

that models relying on machine learning had better prediction power than traditional statistical techniques based on general linear model (GLM) [14, 15, 16, 17]. It is still necessary to develop machine learning-based prediction models and compare their prediction power with the prediction power of GLM-based regression models for proving the usefulness of machine learning in the medical field. This study developed a model for predicting healthy hearing people's speech acceptability for children with cochlear implants using multiple regression analysis, support vector regression, and random forest and evaluated the prediction performance of the model by comparing mean absolute errors and root mean squared errors.

II. Research Methods

A. Study Subjects

It is a descriptive study that identified the factors associated with the speech acceptability for children with cochlear implants perceived by people with healthy hearing and this study targeted 91 hearing-impaired children between four and eight years old who resided in Seoul, Incheon, and Suwon and had worn cochlear implants at least one year and less than five years. The subjects of this study were the same as Byeon [5]. The study subjects were (1) the hearing-impaired who had worn cochlear implants at least one year, (2) those who received hearing rehabilitation regularly after surgery, and (3) those who were using oral speech during a conversation. This study excluded children with a cognitive disorder, an affection disorder, visual impairment, Autism spectrum, and development disabilities in addition to hearing impairment. The power was tested using G-Power version 3.1.9.7 (Universität Mannheim, Mannheim, Germany) (Fig. 1), and the minimum sample size was derived as 80 samples when the number of predictors was seven, the significance level was a=0.05, power (1-B) was 0.8, and the effect size (f2) was 0.2. This study's sample size satisfied the appropriate sample size for testing the statistical significance (Fig. 2).

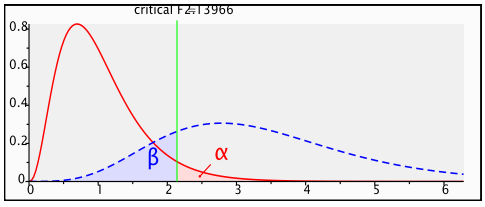


Fig. 1. Result of Power Analysis.

F tests – Linear multiple regression: Fixed model, R^2 deviation from zero			
Analysis: A priori: Compute required sample size			
Input:	Effect size f²	=	0.2
	α err prob	=	0.05
	Power (1-β err prob)	=	0.8
	Number of predictors	=	7
Output:	Noncentrality parameter λ	=	16.0000000
	Critical F =	2.1396555	
	Numerator df	=	7
	Denominator df	=	72
	Total sample size	=	80
	Actual power	=	0.8061255

Fig. 2. Results of Sample Size Calculation.

B. Measurement

Speech data of children wearing cochlear implants (CI) were collected through two tasks, speaking and reading. The reading sentence was "Once upon a time, there was a young tiger living in a village. The young tiger was very curious" referring to Yoon [18]. The speaking task was to introduce oneself in the form of "Nice to meet you. My name is OOO." It was recorded using the Multi-Dimensional Voice Program (MDVP, Key Pentax, USA) installed on the computer in a quiet room without noise, and the microphone (Shure BETA58A) used for recording was located 10cm below the child's mouth.

The outcome variable, healthy hearing people's speech acceptability for children wearing CI, was evaluated by 80 college students (freshman and sophomore) who did not have prior knowledge of children with a cochlear implant. Each evaluator evaluated the speech acceptability of each child after listening to the speech data of the child once, which was played on a computer through a speaker in a noise-free place, and there was a 5-second interval between speech data. Speech acceptability was measured using a visual analog scale. The evaluators indicated the degree of speech acceptability perceived by them on a 100mm straight line where 0 was marked as "impossible to understand" and 100 was marked as "fully understandable" [19]. After the first evaluation was completed, the second evaluation was performed by changing the presentation order of the speech data. The mean values of the first and second evaluations were defined as the final scores of speech acceptability for the subjects' reading and speaking.

Explanatory variables included gender, age, household income, the period of wearing cochlear implants, corrected hearing, auditory-language rehabilitation period, pitch, loudness, and quality. Corrected hearing was defined as the mean threshold decibels (dB) of hearing tests measured in the ranges of 250, 500, 1k, 2k, and 4kHz after wearing a cochlear implant. Where the subject wore cochlear implants for both ears, the mean threshold value was used. When a cochlear implant was used for one ear and a hearing aid was used for the other ear, only the hearing of the cochlear implant side was used. Pitch, loudness, and quality were defined by analyzing the speech data recorded in MDVP.

C. Analysis Methods

This study developed a model for predicting the speech acceptability of children wearing CI using multiple regression analysis, support vector regression analysis, and random forest algorithm. This study also evaluated and validated the model to test the prediction performance of the developed model. This study randomly divided the data into a training dataset (70%) and a test dataset (30%) for validating the prediction performance; the training dataset was used to develop a prediction model and the test dataset was used to evaluate the prediction performance (mean absolute error and root mean squared error) by using the test dataset. All analyses were performed using R version 4.0.2 (Foundation for Statistical Computing, Vienna, Austria) and Python version 3.8.0 (https://www.python.org). The schematic diagram of the study is presented in Fig. 3.

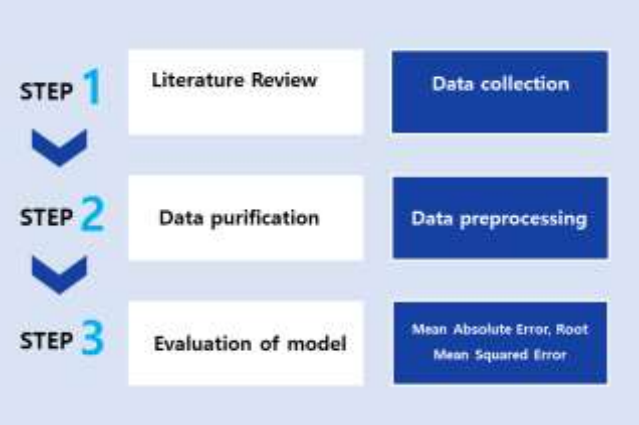


Fig. 3. The Schematic Diagram of the Study.

D. Multiple Regression Analysis

Multiple regression analysis is an analysis that models the relationship between data while reiterating the process to minimize the error between the given data and the values obtained by the selected learning model. Linear regression is a method of analyzing the linear relationship between a dependent variable and at least one independent variable. When the dependent variable is a continuous variable, it can be analyzed using linear regression. When using the multiple linear regression analysis, it is possible to identify the influence (weight) of each independent variable on the dependent variable by estimating the regression coefficient. The least squares method or the maximum likelihood estimation method is used to estimate regression coefficients when modeling using the multiple linear regression method to predict results. Generally, the least squares method is used to make a regression model and analysis prediction results. The least squares method uses a method that minimizes the error of the model (the difference between estimated values of a model and actual observations) for estimate regression coefficients. Therefore, it searches for a model that can estimate values that are close to the actual results. This study also constructed a multiple linear regression model by applying the least squares method. An example of the least squares method is presented in Fig. 4.

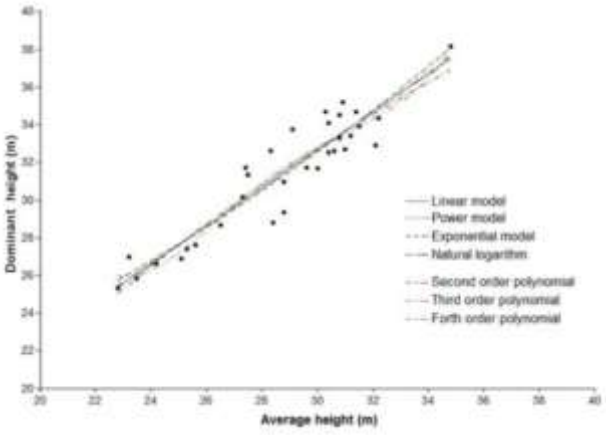


Fig. 4. Examples of Regression Analysis by Least Squares Method with different Model Formulations [20].

E. Random Forest

Random forest is composed of multiple decision trees. The goal of random forest is to make more accurate predictions by making multiple decision tree models. Random forest is a decision tree-based ensemble method, which generates numerous random samples through a bootstrap method that randomly extracts samples with replacement of the same sample size from the training dataset, learns an independent decision tree for each sample dataset, and determines the final model by summarizing the results. The ensemble method is to create a final prediction model by generating multiple prediction models from a given data and then combining them. Many previous studies have shown that the ensemble method can improve the predictive power of the model [21,22]. Moreover, random forest has smaller prediction errors with more decision trees and it does not overfit even if there are many decisions, which are advantages of random forest. The concept of random forest is presented in Fig. 5.

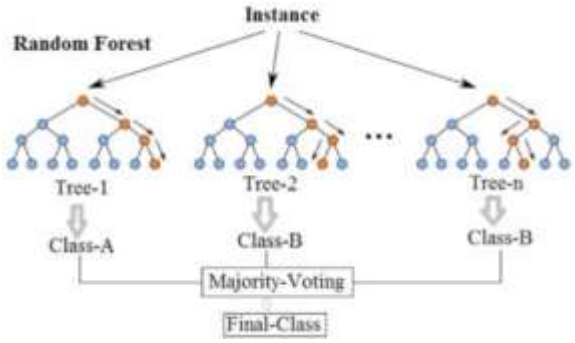


Fig. 5. The Concept of Random Forest [23].

F. Support Vector Regression Analysis

Support vector regression is a regression model based on a support vector machine (SVM). SVM finds the optimal hyperplane that classifies the data into the most suitable classes by maximizing the margin for classifying input data by expressing the data in a high-dimensional vector space using a kernel function. Support vector regression is an extension of this SVM so that SMV can be applied to regression analysis. It is used to predict a random error tolerance value by introducing an e-insensitive loss function [24]. Support vector regression, like SVM, uses a kernel function to converting training data into points in feature space and then performs learning in feature likelihood. However, SVM and support vector regression are different in the aspect that SVM is a machine learning to classify "+1 class" and "-1 class", while support vector regression is a method to generalize class for predicting random error tolerance values using a regression function [25]. Support vector regression has the advantage of having high explanatory power even for data showing nonlinearity or complex patterns. However, it takes a long time to learn because computational complexity is high and it is not easy to interpret the model because it is not possible to analyze the direct relationship between independent and dependent variables, which are disadvantages. Moreover, support vector regression converts a nonlinear feature dimension, which cannot be linearly separated linearly, into a high-dimensional linear regression problem using a kernel function for nonlinear

expansion. The kernel function generally used in this process is a linear, polynomial, or radial basis function. The concept of support vector regression is presented in Fig. 6.

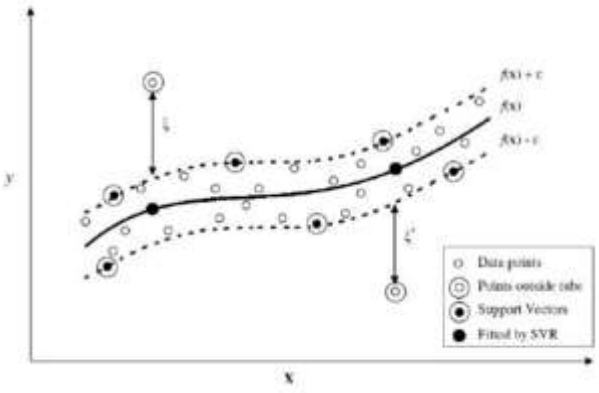


Fig. 6. The Concept of Support Vector Regression [26].

G. Evaluating the Prediction Performance of the Model

A multiple linear regression model was built by using a regression coefficient estimation method based on the least squares method. While conducting random forest analysis, the limit of decision tree development was set to 100. Support vector regression was analyzed using a linear kernel function, the most basic kernel function, c (the parameter determining the generalization of a regression model) was 15.0, and the e-insensitive loss function (a precision parameter) was set as 0.001. This study compared mean absolute errors and root mean squared errors to evaluate the prediction performance of the developed models. Since random forest includes randomness, the model was developed while fixing the seed (#123456) during repeated measurements.

III. Results

A. Comparing the Performance of Models for Predicting Healthy hearing People's Speech Acceptability for Children Wearing CI

Table I shows the mean absolute errors and root mean squared errors of the speech acceptability prediction model for children wearing CI using multiple regression analysis, support vector regression analysis, and random forest. This study defined that a model with the smallest mean absolute error and root mean squared error was the best model with the best prediction performance. The results of this study showed that the random forest algorithm (mean absolute errors=0.81and root mean squared error=0.108) was the best model for predicting the speech acceptability of children wearing CI.

B. The Importance of Variables in the Final Model (Random Forest) for Predicting the Speech Acceptability for Children with CI Wearers

The normalized importance of random forest variables (the final model) is presented in Fig. 7. It was found that pitch, loudness, quality, the duration of wearing cochlear implants, the duration of aural rehabilitation, corrected hearing, and age were major variables with high weight in predicting the speech acceptability of children wearing CI. Among these variables, pitch was the most important factor in the final model.

TABLE I. Mean Absolute Errors and Root Mean Squared Errors of Models for Predicting the Speech Acceptability for Children Wearing CI using Multiple Regression Analysis, Support Vector Regression, and Random Forest

Type of model	Mean absolute error	Root mean squared error
Multiple regression	0.084	0.113
Support vector regression	0.081	0.109
Random forest	0.081	0.108

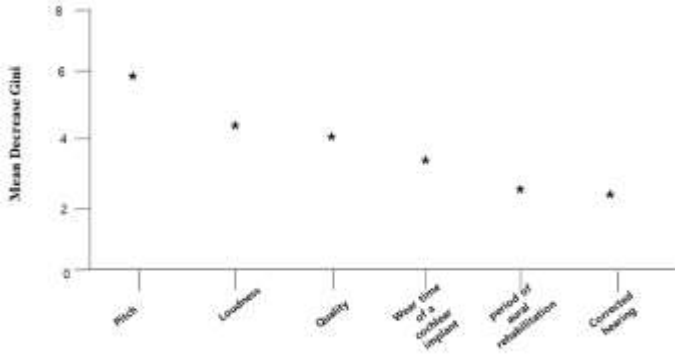


Fig. 7. The Normalized Importance of Variables in the Model for Predicting Healthy Hearing People's Speech Acceptability for Children Wearing CI based on Random Forest (Only the Results of the Top 6 Variables are Presented).

IV. Conclusion

This study developed a model for predicting healthy hearing people's speech acceptability for children wearing CI and found that pitch, loudness, and quality were main variables with higher weight for predicting the speech acceptability of children wearing CI. Among them, pitch was the most important factor in the final model. Factors affecting speech acceptability can be divided into segmental factors such as the errors in individual consonants and vowels and supra-segmental factors such as stress, speaking rate, voice quality, and intensity. Dagenais et al. [27] evaluated dysarthria and showed that speech acceptability was significantly correlated with speaking rate. Moreover, Lee et al. [19] analyzed the speech acceptability of hearing-impaired adults and reported that the speech acceptability of them was more strongly correlated with supra-segmental factors than segmental factors, and consonant accuracy, intonation, resonance, and speech rate were major variables influencing speech acceptability. Previous studies [28, 29] that analyzed the acoustic and phonetic characteristics of speech made by hearing impaired children with wearing CI showed that the pitch and quality related indices of children wearing CI were different from those of healthy hearing children. Hsu et al. [30], who evaluated auditory senses, also showed that the speech characteristics of children wearing CI were different from those of healthy hearing children in terms of pitch, quality, and resonance. In summary, the results of this study suggested that the speech characteristics of hearing impaired children with wearing CI, which the listener felt unnatural, were mostly due to acoustic-phonetic characteristics such as pitch and loudness among various speech-related factors such as age and gender.

Another finding of this study was random forest had the best prediction performance among multiple regression

analysis, support vector regression analysis, and random forest after comparing the accuracy of the models for predicting the healthy hearing people's speech acceptability for children wearing CI. This study developed prediction models using random forest (a machine learning technique), support vector regression analysis (a machine learning technique), and multiple regression analysis (a GLM analysis technique) and evaluated prediction performance by calculating mean absolute errors and root mean squared errors. The results of this study showed that random forest-based speech acceptability prediction model for children wearing CI showed the smallest mean absolute error and root mean squared error among the three models. This result agreed with the results of previous studies [14, 15, 16, 17] indicating that random forest-based models performed better than regression models in predicting diseases. The results of this study support the possibility that the accuracy of the ensemble model may be better than that of GLM. Furthermore, they imply that the predictive performance of random forest will be the best among ensemble models when developing a machine learning model using speech data of children wearing CI. Further studies are needed to prove the prediction performance of random forest by comparing accuracy using data from various fields.

ACKNOWLEDGMENT

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

REFERENCES

[1] M. Castro-Neto, Y. Jeong, M. K. Jeong, and L. D. Han, AADT prediction using support vector regression with data-dependent parameters. Expert Systems with Applications, vol. 36, no. 2, pp. 2979-2986, 2009.

[2] M. A. Svirsky, A. M. Robbins, K. I. Kirk, D. B. Pisoni, and R. T. Miyamoto, Language development in profoundly deaf children with cochlear implants. Psychological Science, vol. 11, no. 2, pp. 153-158, 2000.

[3] H. Byeon, Development trends of online-based aural rehabilitation programs for children with cochlear implant coping with the fourth industrial revolution and implication in speech-language pathology. International Journal of Advanced Computer Science and Applications, vol. 10, no. 10, pp. 25-31, 2019.

[4] H. Bruijnzeel, F. Ziylan, I. Stegeman, V. Topsakal, and W. Grolman, Systematic review to define the speech and language benefit of early. Audiology and Neurotology, vol. 21, no. 2, pp. 113-126, 2016.

[5] H. Byeon, Developing a model for predicting the speech intelligibility of South Korean children with cochlear implantation using a random forest algorithm. International Journal of Advanced Computer Science and Applications, vol. 9, no. 11, pp. 88-93, 2018.

[6] M. C. Allen, T. P. Nikolopoulos, and G. M. O'Donoghue, Speech intelligibility in children after cochlear implanation. Otology & Neurotology, vol. 19, no. 6, pp. 742-746, 1998.

[7] A. E. Geers, Speech, language, and reading skills after early cochlear implantation. Archives of Otolaryngology–Head & Neck Surgery, vol. 130, no. 5, pp. 634-638, 2004.

[8] T. L. Whitehill, Assessing intelligibility in speakers with cleft palate: a critical review of the literature. The Cleft Palate-Craniofacial Journal, vol. 39, no. 1, pp. 50-58, 2002.

[9] L. W. Ellis, Magnitude estimation scaling judgments of speech intelligibility and speech acceptability. Perceptual and Motor Skills, vol. 88, no. 2, pp. 625-630, 1999.

[10] I. K. Y. Law, E. P. M. Ma, and E. M. L. Yiu, Speech intelligibility, acceptability, and communication-related quality of life in Chinese alaryngeal speakers. Archives of Otolaryngology–Head & Neck Surgery, vol. 135, no. 7, pp. 704-711, 2009.

[11] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, Disease prediction by machine learning over big data from healthcare communities. Ieee Access, vol. 5, pp. 8869-8879, 2017.

[12] H. Byeon, Best early-onset Parkinson dementia predictor using ensemble learning among Parkinson's symptoms, rapid eye movement sleep disorder, and neuropsychological profile. World Journal of Psychiatry, vol. 10, no. 11, pp. 245-259, 2020.

[13] Y. Lee, and J. Kim, Artificial intelligence technology trends and IBM Watson references in the medical field. Korean Medical Education Review, vol. 18, no. 2, pp. 51-57, 2016.

[14] H. Byeon, Development of a depression in Parkinson's disease prediction model using machine learning. World Journal of Psychiatry, vol. 10, no. 10, pp. 234-244, 2020.

[15] M. Khalilia, S. Chakraborty, and M. Popescu, Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making, vol. 11, no. 1, pp. 51, 2011.

[16] H. Byeon, Exploring the predictors of rapid eye movement sleep behavior disorder for Parkinson's disease patients using classifier ensemble. Healthcare, vol. 8, no. 2, pp. 121, 2020.

[17] H. Byeon, Application of machine learning technique to distinguish Parkinson's disease dementia and Alzheimer's dementia: predictive power of Parkinson's disease-related non-motor symptoms and neuropsychological Profile. Journal of Personalized Medicine, vol. 10, no. 2, pp. 31, 2020.

[18] M. S. Yoon, Variables for predicting speech acceptability of children with cochlear implants. Phonetics and Speech Sciences, vol. 6, no. 4, pp. 171-179, 2014.

[19] S. E. Lee, H. H. Sim, C. M. Nam, J. Y. Choi, and E. S. Park, Auditory-perceptual evaluation of the speech of hearing-impaired adults: based on suprasegmental factors, speech intelligibility, and speech acceptability. Communication Sciences & Disorders, vol. 15, no. 4, pp. 477-493. 2010.

[20] A. Ferezliev, R. Mavrevski, and A. Delkov, Correlation between average and dominant height of middle-aged douglas fir plantations in the North-West Rhodopes. Silva Balc, vol. 19, no. 2, pp. 13-26, 2018.

[21] P. T. Noi, and M. Kappas, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors, no. 18, no. 1, pp. 18, 2018.

[22] H. Byeon, Is the random forest algorithm suitable for predicting Parkinson's disease with mild cognitive impairment out of Parkinson's disease with normal cognition?. International Journal of Environmental Research and Public Health, vol. 17, no. 7, pp. 2594, 2020.

[23] A. Chakure, Random Forest Classification. The Stratup, 2019.

[24] A. J. Smola, and B. Schölkopf, A tutorial on support vector regression. Statistics and computing, vol. 14, no. 3, pp. 199-222, 2004.

[25] M. Castro-Neto, Y. Jeong, M. K. Jeong, and L. D. Han, AADT prediction using support vector regression with data-dependent parameters. Expert Systems with Applications, vol. 36, no. 2, pp. 2979-2986, 2009.

[26] S. K. Lahiri, and K. C. Ghanta, The support vector regression with the parameter tuning assisted by a differential evolution technique: Study of the critical velocity of a slurry flow in a pipeline. Chemical Industry and Chemical Engineering Quarterly, vol. 14, no. 3, pp. 191-203, 2008.

[27] P. A. Dagenais, G. R. Brown, and R. E. Moore, Speech rate effects upon intelligibility and acceptability of dysarthric speech. Clinical Linguistics & Phonetics, vol. 20, no. 2/3, pp. 141-148, 2006.

[28] M. S. Yoon, E. A. Choi, and Y. J. Sung, A Comparison of voice analysis of children with cochlear implant and with normal hearing. Journal of the Korean Society of Speech Sciences, vol. 5, no. 4, pp. 71-78, 2013a.

[29] L. H. P. Nguyen, J. Allegro, A. Low, B. Papsin, and P. Campisi, Effect of cochlear implantation on nsality in children. Ear, Nose & Throat Journal, vol. 87, no. 3, pp. 138-143, 2008.

[30] H. W. Hsu, T. J. Fang, L. A. Lee, Y. T. Tsou, and S. H. Chen, Multidimensional evaluation of vocal quality in children with cochlear implants: a cross-sectional, case-controlled study. Clinical Otolaryngology, vol. 39, no. 1, pp 32-38, 2014.

Advanced Debugger for Arduino

Jan Dolinay[1], Petr Dostálek[2], Vladimír Vašek[3]
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlín, Czech Republic

***Abstract*—This article describes improved version of our source-level debugger for Arduino. The debugger can be used to debug Arduino programs using GNU debugger GDB with Eclipse or Visual Studio Code as the visual front-end. It supports all the functionally expected from a debugger such as stepping through the code, setting breakpoints, or viewing and modifying variables. These features are otherwise not available for the popular AVR-based Arduino boards without an external debug probe and modification of the board. With the presented debugger it is only needed to add a program library to the user program and optionally replace the bootloader. The debugger can speed up program development and make the Arduino platform even more usable as a tool for controlling various experimental apparatus or teaching computer programming. The article focuses on the new features and improvements we made in the debugger since its introduction in 2016. The most important improvement over the old version is the support for inserting breakpoints into program memory which allows debugging without affecting the speed of the debugged program and inserting breakpoints into interrupt service routines. Further enhancements include loading the program via the debugger and newly added support for Arduino Mega boards.**

Keywords*—Arduino; debugger; microcontroller; software debugging

I. Introduction

Arduino is a very popular prototyping platform with a microcontroller (MCU). It started as an educational tool in 2003 and evolved into a widespread platform for prototyping, controlling various devices and for teaching computer programming. It is now frequently used in courses focused on embedded systems, robotics, and the like. For example, [1] describes successful use of the platform in computer science capstone course, [2] used Arduino-based custom board to increase student's interest in programming courses and [3] utilized Arduino as the base for their educational mobile robot. A comprehensive review on this topic was presented e.g., by [4]. Arduino is also used in scientific laboratories as a low-cost multipurpose device for controlling various experimental apparatus in a wide range of areas. For example, [5] concludes that Arduino boards may be inexpensive tool for many psychological and neurophysiological labs, [6] based their device to abate tremors for patients with Parkinson's disease on this platform, [7] uses Arduino to generate pulsatile flow rate for biofluid dynamics research, [8] uses distributed network of Arduino boards acting as remote servers in a system controlling capacitive energy storage and [9] used Arduino for real-time monitoring of air quality in urban area. Yet another area of its use is the emerging technology of Internet of Things (IoT), as described by [10, 11] and others.

The Arduino hardware is a printed circuit board with a microcontroller. A program must be written, built, and uploaded to the MCU for the Arduino to be able to perform requested tasks. There is a software tool, integrated development environment (IDE), provided with the platform to accomplish this. It is also possible to use other tools to create the programs for Arduino, for example, Eclipse or Visual Studio Code. These tools offer more functionality than the simplistic Arduino IDE and are therefore preferred by many advanced users.

One feature which is commonly missed by advanced users is a source-level debugger. The Arduino IDE does not provide any interface for source-level debugging. The alternative IDEs do provide such interface, but the most popular Arduino boards based on AVR microcontrollers, such as Uno, Mega or Nano, cannot be debugged without an external debug probe and alteration of the board. Thus, most users debug their programs by printing textual messages to serial interface, which is usable, but not very effective or comfortable.

A debugger that lets the user stop the program, view, and modify variables or execute the program step by step can save significant amount of time in localizing problems, especially in more complex projects. It can also greatly improve the usability of the Arduino platform for teaching computer programming. For the novice programmers it is easier to create the mental model of the programming constructs they are learning if they can use a debugger to single step through lines of code, set breakpoints, and watch the internal state of the program as well as the outside effects, such as an LED turning on. Debugger is also an essential tool for teaching debugging skills, which is recognized as an important part of computing curricula [12, 13].

As follows from the above a source level debugger is a highly desirable feature, which can make the Arduino platform more usable both for teaching programming and for implementing various devices. We developed first version of such a debugger in 2016 [14]. This debugger had several limitations which affected the performance of the debugged program and thus the usability of the tool. In this article, we present a significantly improved version of the debugger, which overcomes limitations of the first version and offers features and comfort of use at the level expected from fully-fledged, hardware-based debugger.

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and also by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03 xxx.

II. Debugger Performance Improvements

As already mentioned, in 2016 we developed a source-level debugger for the AVR-based Arduino boards. The unique feature of this debugger is that it makes it possible to debug programs at source level without any external tools. This can be particularly useful for educational purposes – there is no extra cost of buying a hardware debug probe for each workplace in the lab and no extra work with modifying the Arduino boards to be able to communicate with the probe.

However, the first version of the debugger had several limitations - it was implemented for Arduino Uno only, the program execution was significantly slower when any breakpoints were set, and it was not possible to set breakpoints into interrupt service routines (ISR). After receiving positive feedback from the users, we decided to improve the debugger to overcome these limitations.

A. Debugging Options for Arduino

There are three options for debugging the AVR-based Arduinos besides printing messages to serial line. First option is to use a hardware debugger – a debug probe, which connects to the debug pin of the MCU. Unfortunately, there is a capacitor attached to this pin on the Arduino boards which must be disconnected for the communication with the debug probe to work. Newer revisions of the board are equipped with a solder bridge which can be cut to enable this feature, making the modification relatively easy. Nevertheless, it is a modification which needs to be reverted if the normal program uploading via bootloader is needed. Another problem is that the debug protocol is proprietary and therefore a commercial debug probe must be obtained. The prices of such probes start at around $50 for the most affordable Atmel-ICE-PCBA tool.

The second option is to use VisualMicro Arduino IDE for Visual Studio, which contains a tool called Serial debugger. This debugger is based on inserting code into the user program to communicate with the IDE. This added code is not visible to the user and the user experience is quite satisfactory, but there are major limitations to this approach - it is not possible to insert breakpoints during debugging; a rebuild and re-upload is required. Also, stepping through the code is not possible - the program can only be run from one breakpoint to the next one. Moreover, the VisualMicro is a paid software; the prices start at $12 for a one-year student license or $49 for a permanent license. All this, together with the fact that it is only available as an extension for Visual Studio, a complex and hardware intensive IDE, probably makes it less than ideal solution for many users.

The third option is a debugger based on GDB stub mechanism, which is described in the next section. The advantage of this approach is that the features of such a debugger are very similar to a hardware debugger, including stepping through the code, inserting breakpoints in runtime, and viewing and modifying the variables. Also, the stub is not limited to certain IDE; it can be used with GDB in command line as well as with various IDEs. We verified the solution with Eclipse IDE, Visual Studio Code and PlatformIO which are all free, multiplatform, and relatively light-weight environments. As far as we know, our GDB stub presented here is the only such stub for the AVR based Arduinos available.

To summarize, to be able to debug Arduino programs at source level, the other options besides our solution are either modifying the board and buying a debug probe for approx. $50 or buying the Visual Micro software solution which has limited features. We believe that our solution is an attractive option especially for educators, as it is completely free, can be used in Windows, Linux or Mac and works with the Arduino board as-is, without any hardware modification.

B. Principle of Operation

The debugger is based on so-called debugger stub for the GNU debugger GDB. Debugger stub is a small program that runs on the debugged computer and communicates with the debugger running on development (host) computer [15].

To enable debugging, users must insert a program library (the stub) into their code and then they are able to connect to the running program and debug it with the GDB. We first presented this solution with Eclipse IDE, but it can also be used from command line or with any IDE that can integrate with GDB, notably the popular open-source editor Visual Studio Code. The principle of communication is shown in Fig. 1. Our software component, GDB stub, is part of the user program running on the Arduino board (target system). The stub handles serial communication with the GDB debugger running on the host system (desktop computer). This way the GBD can control the program, view the memory, etc.

C. New Breakpoints Implementation

The key new feature described in this paper is the ability to set breakpoints into program memory. In general, breakpoints are implemented by modifying the program code at the position where the execution should be suspended. The original instruction is replaced by another instruction that redirects the execution to the debugger. This is easily achieved on desktop computers, as the program is located in RAM memory, which it is easy to modify. The problem with using this technique in microcontrollers is that the program memory is typically based on flash technology and cannot be easily rewritten in runtime.

MCUs are commonly equipped with a debug module which takes care of inserting the breakpoints, single stepping etc. However, to access this module a special hardware - a debug probe is required. Moreover, the AVR-based Arduino boards can only be used with such a probe after modification of the printed circuit board, as mentioned in Section A.

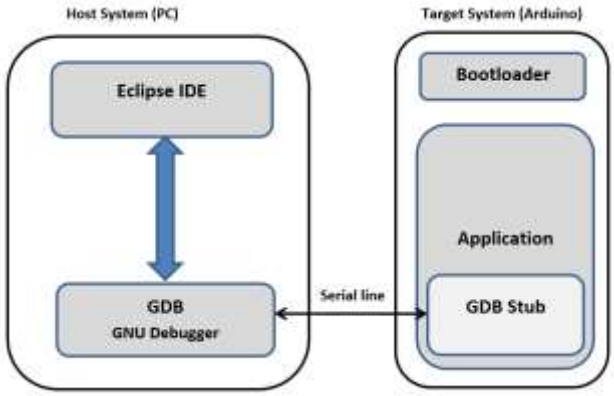


Fig. 1. Principle of Debugging with our GDB Stub.

We aimed to provide accessible debugging option without the relatively expensive hardware probe and modification of the board. That is why we implemented the debugger using the debugger stub technique mentioned above, without using the debug module of the MCU. The first version of our debugger introduced in [14] did not insert breakpoints into the program memory because it seemed too complicated at that time. Instead, breakpoints were implemented by comparing the current position in the program with a list of desired breakpoint addresses after executing every instruction of the CPU. Obviously, this considerably decreases the execution speed of the debugged program, but in many situations, it is not a problem and the debugger can be successfully used. The advantage of this approach is that the users simply add a library to their program; no other action is required.

However, setting breakpoint directly to the program memory promises significant benefits and we decided to implement this feature in the new version. In the following text we refer to these new breakpoints as "flash breakpoints" and to the older breakpoints as "RAM breakpoints". The advantage of the flash breakpoints is that the program can run at full speed, with virtually no intrusion, until a breakpoint is hit. Moreover, flash breakpoints can be set into interrupt service routines (ISR), which is not possible with the RAM breakpoints. This follows from the principle of operation of the RAM breakpoints – after each instruction of the main program an ISR must be executed to examine current position of the program - and the AVR CPU can only execute one ISR (which is used by the debugger) and the main program in this mode. On the other hand, flash breakpoints are implemented by replacing the original instruction at the position where the program should stop with another instruction that redirects execution back to the debugger. Thus, they do not require any use of an interrupt for monitoring current position of the program which would grade the execution speed. The details of the implementation are provided in the following section.

III. Results and Discussion

In this section we describe the implementation of the new version of the debugger stub which can be useful for better understanding of the features and limitations of this solution. We also present sample cases of running the debugger.

A. New Breakpoints and Program Load Implementation

To implement the flash breakpoints, we needed to solve two problems. First problem is that on the ATmega328 MCU it is only possible to rewrite the flash memory from code running in special memory section called the bootloader section. This section already contains the Arduino bootloader which handles loading user programs from the IDE. In normal course of operation, user programs (including our debugger stub) cannot be loaded into this section.

To solve this problem, we developed custom version of the bootloader which works in the same way as the standard bootloader - it allows uploading the user program without debugging, but additionally it provides service to the debugger stub to write to program memory. The principle is depicted in Fig. 2. When the stub needs to set a breakpoint in the program memory it calls a routine in the custom bootloader to modify the flash memory. The bootloader also provides a routine for loading new application program as described later.

The second problem was how to replace the original instruction in the memory when setting a breakpoint. To stop the program, we need to execute some code that will pass the control from the debugged program to the debugger. Often, there is a special instruction in the instruction set of the target CPU to break the execution and jump to the debugger, or an instruction for software interrupt which can be used to pass the control to an appropriate ISR handled by the debugger. However, in the AVR architecture neither of these are available.

The simple solution would seem to be to replace the original instruction with a jump to the debugger, but such a jump would require overwriting several bytes of the memory – the jump instruction together with its target address. Yet we can only replace single instruction by the breakpoint; we cannot overwrite the following instruction. Consider the case of setting a breakpoint one instruction before the location which is the target of a jump or a subroutine call. If the breakpoint replaces not only the intended instruction but also the following one, the program will crash when it jumps to the now-damaged location after the breakpoint. From this it follows that the instruction to be used as a breakpoint must not be longer than the shortest instruction of the CPU – which is just one word. This significantly limits the available options.

Our first solution was to use an external interrupt which would be asserted all the time but disabled in the peripheral, and to replace the original instruction with an instruction to enable the interrupt. Such an instruction fits into single word but as we found out the execution of the program does not stop immediately at the instruction which enables the interrupt; the program only stops at the next instruction. This would be unacceptable and thus this solution had to be abandoned.

The working solution proved to be using a relative jump instruction (RJMP) with -1 as the target address, which is a jump to itself (an endless loop). Thus, the program stops in an infinite loop at the position of the breakpoint. To pass the control to the debugger we use periodic interrupt that checks whether the program is currently at the address of a breakpoint and if so, it calls the debugger code. The watchdog peripheral is utilized to generate the periodic interrupt leaving all the timers free for use by the Arduino software.

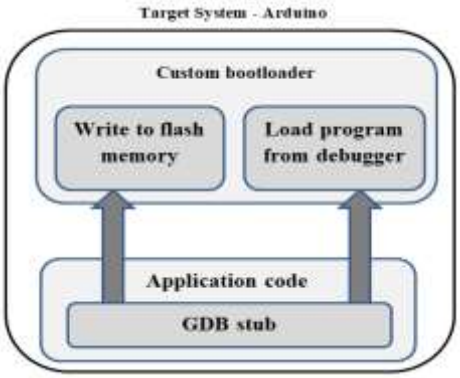


Fig. 2. Interaction of the GDB Stub with the Bootloader.

As follows from the above, to use the flash breakpoints the users need to replace the default bootloader in their Arduino board with our bootloader. Replacing the bootloader is relatively easy and the procedure is well documented by the Arduino community. However, should it be a problem, the debugger can still be used without replacing the bootloader, with RAM breakpoints only. This mode is supported as a compile-time option in the code. It has two advantages – it is readily available for Arduino boards without any modification, and it does not cause wear of the flash memory, as discussed in the next section. In some use-cases, such as school lab for a basic programing course, the RAM-breakpoints mode may be sufficient and preferred.

On the other hand, if the bootloader is replaced and thus writing to flash memory from the user program is enabled, the users can take advantage of another new feature we implemented – the support for loading the program via the debugger. Without this feature the typical workflow for debugging a program is as follows:

- Build the program.
- Upload the program (through the Arduino bootloader).
- Attach the debugger and debug.

With the load support it is possible to upload the program and start debugging with a single click in the IDE. Our debugger stub takes care of receiving the new program and writing it to the program memory of the MCU.

B. Flash Memory Wear Considerations

As already mentioned, the flash breakpoints are implemented by replacing one instruction in the program by a code which transfers the control to the debugger stub. Thus, setting a breakpoint requires overwriting part of the program memory of the MCU. This memory is based on flash technology, which can only endure certain number of write cycles. For the ATmega328 MCU the manufacturer guarantees flash endurance of ten thousand cycles. Although the number is high, memory wear should be considered when using the debugger. Let us briefly discuss this topic.

From user's perspective there are two debugger commands to control execution of the program – a Step command which moves the program to the next line and Continue (Run) command which lets the program run until a breakpoint is hit.

To perform the Step command, the debugger must execute one or more CPU instructions – consider that one line of code in C language may correspond to several CPU instructions. We implement the Step command in the same way as RAM breakpoints – one instruction is executed and then an interrupt is triggered to check whether desired position in the program has been reached. Consequently, the flash memory is not rewritten during step commands.

The Continue command requires writing a breakpoint to program memory - the program should run until a breakpoint is hit. Besides user-defined breakpoints there are also some breakpoints inserted automatically by the debugger. For example, when stepping over a function the GDB places temporary breakpoint at the next instruction after the function call. GDB also removes all breakpoints when the program stops on a breakpoint so that the user can see the stopped program with the original instructions in place and so that the original instruction can be executed when continuing from a breakpoint. When the user resumes the program, all active breakpoints must be written back to memory because the debugger does not know which one will be hit next.

To reduce the memory wear, we implemented a simple optimization so that the breakpoints are written and removed only if it is necessary. When our stub receives command from GDB to remove a breakpoint it notes this request but does not remove the breakpoint (rewrites the flash memory) until the continue command is received. In many cases the breakpoint is removed and then replaced by GDB, but the stub leaves the breakpoint in place thus saving two flash write cycles. Even with this optimization one should keep in mind that the flash memory is overwritten often. It is possible to analyze the number of writes if a compile-time option is enabled in the code of the stub; then there is a global variable which tracks the number of writes to flash memory.

C. Running the Debugger

In the following sections we show the usage of the debugger with focus on the new features. For detailed explanation of the basic setup and usage please refer to our earlier article [14]. The results presented here were obtained on a Window 10 desktop computer with Eclipse Oxygen IDE, 64-bit, version 4.7.3a.

We assume an Arduino Uno board in the original state as it left the factory. First step is to replace the bootloader. For this an in-circuit AVR programmer (ISP programmer) is required. Such programmers are available in many variants for low prices. It is also possible to use another Arduino board as a programmer. The modified bootloader can be found in the source package as a .hex file. This file needs to be loaded into the MCU memory instead of the original bootloader and so-called fuses need to be changed to take into account different size of the new bootloader – the fuse BOOTSZ (size of the bootloader region of the MCU) needs to be set to 1024 words, which means the bootloader occupies 2 kB of memory. After replacing the bootloader, we are ready to use the new features.

D. Building the Program

We will use the typical introductory program which blinks the on-board LED on Arduino pin 13. The user first needs to set up the Eclipse IDE to be able to develop programs for Arduino. The procedure is described in the documentation provided with the source code.

Once the Eclipse project is set up to build the blink program, we can add the code of our debugger stub. This code is located in four files in the avr8-stub folder: avr8-stub.c, avr8-stub.h, app_api.c and app_api.h. The two app_api files provide communication with the bootloader and are only needed if the flash breakpoints or load-via-debugger options are enabled.

As the next step we configure the debugger stub. The constant AVR8_BREAKPOINT_MODE determines whether the flash breakpoints should be used. Default value 1 results in using RAM breakpoints only. Changing it to 0 enables the flash breakpoints.

The constant AVR8_LOAD_SUPPORT determines whether it should be possible to load the program via the debugger. We set the value to 1 to enable this feature. Now we can build the program.

E. Debugging the Program

Once the program is built and an Eclipse debug configuration is created the program can be debugged. Even with the load-via-debugger option enabled we still need to upload the program to the MCU in the standard way (using avrdude tool and bootloader) for the first time because the code of the GDB stub is not yet present in the MCU to be able to receive new programs directly.

After uploading the program to the MCU, we are ready to start debugging. In our earlier article we described using a TCP-to-serial converter on Windows 7 to connect the GDB and the debugger stub because direct serial connection was unstable. With Windows 10 or Linux systems direct serial connection can be used.

Once the code is uploaded to the MCU we can connect to the running program using the Debug button in the IDE. When the connection is established, we should see the program stopped in the debugger – as shown in Fig. 3. The Debug window at the top shows the call stack. The program is stopped in the loop function. Below, in the source window, the line to be executed next is highlighted – it is a call to the delay function. It is now possible to either step into the function and debug the code inside, or step over and execute the function at once.

There is also a variable "counter" which we can examine or modify. Fig. 4 shows the value of the variable as displayed in the IDE when hoovering cursor over the variable name.

To illustrate the benefit of the flash breakpoints - that they do not slow down the execution of the debugged program, we compare the speed of the program when running with flash breakpoints and with RAM breakpoints. First, we insert a breakpoint into the setup function. It will never be hit because the setup function is only executed once when the program starts. However, as described earlier, the presence of the breakpoint should nevertheless slow down the program when using the RAM breakpoint mode.

With the breakpoint set, we resume the program so that it runs at full speed and measure the period of the blinking of the LED. Using the configuration described above we obtain approximately 2 seconds period, as expected given the two 1000 milliseconds delays in the code. This shows that the program speed is not affected by the presence of the breakpoint.

Now we switch the configuration to RAM breakpoints by setting AVR8_BREAKPOINT_MODE to 1 in avr8-stub.h file. After loading the program and running it without a breakpoint we obtain 2 seconds period as in the previous case. However, after setting the breakpoint into the setup function as in the previous case, we obtain period of 7.2 seconds. This means that the program is slowed down by a factor of nearly 4. If a busy loop is used instead of the timer-based Arduino delay function the program is slowed down even more by a factor of 350.

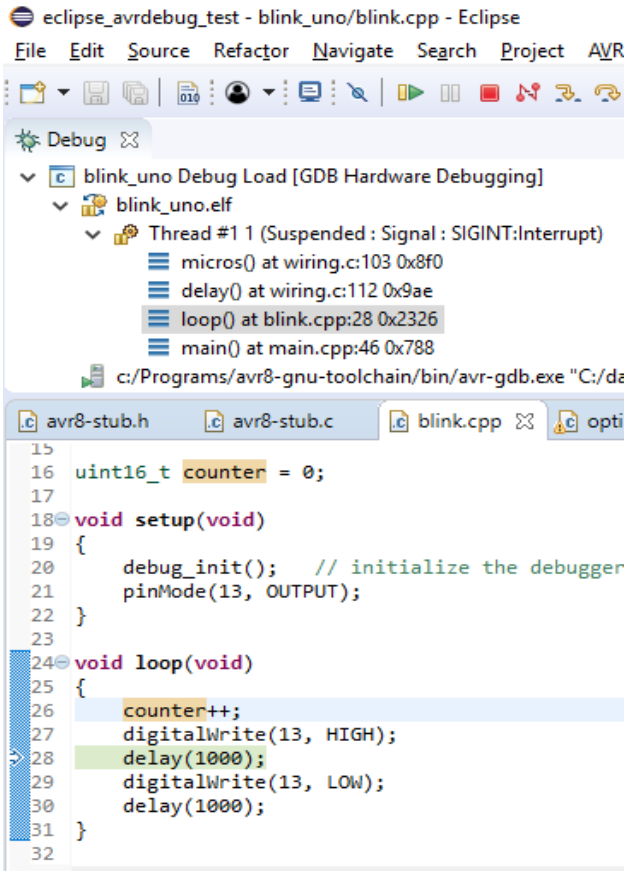


Fig. 3. Sample Program Stopped in the Debugger.

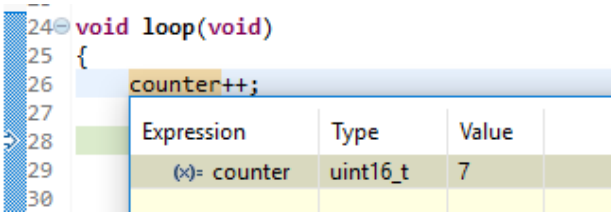


Fig. 4. Viewing Value of a Variable.

F. Loading New Program via the Debugger

With the new feature of loading the program via debugger it is possible to modify and reload the program faster while debugging; just edit the code, click the Debug button and the IDE will save changes, build the program, load it to the MCU and start it in the debugger.

For this to work our debugger stub must be able to receive the program from the IDE and write it to the flash memory of the MCU. To be more precise the GDB debugger issues a load command which our stub supports to load the executable into the target MCU. To enable this feature, the constant AVR8_LOAD_SUPPORT must be defined with value of 1 as described in section D above. Then we need to edit the debug configuration in Eclipse IDE to enable loading the program. This is done by checking the "Load image" box in the Startup tab of the debug configuration. There is detailed description of the procedure in the documentation provided with the source code.

After this we can start the program by clicking the Debug button in the IDE and selecting the debug configuration with load program enabled. To modify the code, we can just stop the program, edit the code, and click the debug button again to upload and debug the modified program.

G. Space Requirements

Table I shows the size of the debugger stub for various configurations of the breakpoints and load support. The exact values will depend on compiler version and configuration, but the presented numbers can be used as estimates of how much space is required to add the debugger support to the program. As can be seen the full-featured version with both breakpoints in flash and load-via-debugger enabled uses about 5.5 kB of program memory. Together with the increased size of the modified bootloader which is required for this configuration (2 kB) adding debugger support to a program requires 7.5 kB out of the total 32 kB of program memory available in the MCU.

H. Use of the Debugger

As already mentioned, presented debugger can be employed without any additional hardware and is free of cost which makes it suitable for use in programming courses which already have the necessary hardware - an Arduino board, and wish to extend the course with introduction to debugging. The improved version of the debugger described here provides better user experience than the old version and allows debugging even time-sensitive code. Besides the educational courses the debugger can also be useful to anyone developing Arduino programs who wants to use a debugger yet is not ready to invest the time and money to switching to a professional development environment.

The current version of the debugger can be used with Arduino boards with the ATmega328 microcontrollers, which includes Uno, Nano and Micro and for the Arduino Mega boards with ATmega2560 and ATmega1280 MCUs.

We are still seeking the ideal development environment to be used with the debugger. The environment based on Eclipse IDE shown above provides complete control of the processes but is quite complicated to set up. A promising alternative seems to be Visual Studio Code editor with Arduino extension which is considerably easier to set up and use for beginner programmers.

TABLE I. CODE AND DATA SIZE FOR DEBUGGER CONFIGURATIONS

Configuration	Program size in bytes	Data size in bytes
Flash breakpoints with load enabled	5446	347
Flash breakpoints with load disabled	5374	307
RAM breakpoints with load enabled	4904	342
RAM breakpoints with load disabled	4658	277

IV. CONCLUSION

In this article we presented new version of the source-level debugger for Arduino. The most important of the new features is the support for writing breakpoints to program memory and loading the program via the debugger. To implement these features a custom version of the Arduino bootloader was created which makes it possible for our debugger stub to write to the program memory. We also had to develop a way to replace the original instruction of the program with a breakpoint to transfer the control to the debugger with the constraint of not overwriting more than one instruction of the original program. Furthermore, to reduce the wear of the flash memory from unnecessary insertion and removal of breakpoints by the GDB debugger, we implemented an algorithm in the code which only rewrites the memory if necessary. Yet another new feature is that the debugger now works also with Arduino Mega boards, which extends its range of use into the area of larger program with many inputs and outputs.

We believe that with these new features the debugger offers user experience similar to a fully-fledged hardware debugger. The advantage of this solution is that, unlike the hardware debug probe, it is free and requires no changes in the Arduino board. It can be helpful in embedded programming courses, student's projects, for controlling lab experiments or in any other project based on the Arduino platform. In future we would like to add support for more Arduino boards and simplify the process of setting up the development environment for debugging. The source code of the debugger stub and detailed instructions for use can be found at https://github.com/jdolinay/avr_debug.

REFERENCES

[1] P. Bender and K. Kussmann, "Arduino based projects in the computer science capstone course", Journal of Computing Sciences in Colleges, Vol. 27, no. 5, pp. 152-157, 2012.

[2] I. Perenc, T. Jaworski and P. Duch, "Teaching programming using dedicated Arduino Educational Board", Comput Appl Eng Educ., vol. 27, no. 4, 943-954, 2019.

[3] F. M. López-Rodríguez and F.J. Cuesta, "Andruino-A1 Low-Cost Educational Mobile Robot Based on Android and Arduino", Journal of Intelligent & Robotic Systems, vol. 81, no. 1, 63-76, 2016.

[4] M. El-Abd. "A Review of Embedded Systems Education in the Arduino Age: Lessons Learned and Future Directions", International Journal of Engineering Pedagogy, vol. 7, no. 2, 79-93, 2017.

[5] A. D'Ausilio, "Arduino, a low-cost multipurpose lab equipment", Behavior Research Methods, vol. 44, no. 2, 305-313, 2012.

[6] J. Hinostroza-Qui˜nones and M. Vasquez-Cunia, "Non-invasive Device to Lessen Tremors in the Hands due to Parkinson's Disease," International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, pp. 735-738, 2020.

[7] M. R. Najjari and M.W. Plesniak, "PID controller design to generate pulsatile flow rate for in vitro experimental studies of physiological flows", Biomedical Engineering Letters, vol. 7, no. 4, 339-344, 2017.

[8] K. I. Mekler, A.V. Burdakov, D.E. Gavrilenko and S. S. Garifov, "A new control system for the capacitive energy storage of the GOL-3 multiple-mirror trap", Instruments and Experimental Techniques, vol. 60, no. 3, 345-350, 2017.

[9] J. Balen, S. Ljepic, K. Lenac and S. Mandzuka, "Air Quality Monitoring Device for Vehicular Ad Hoc Networks: EnvioDev", International Journal of Advanced Computer Science and Applications, vol. 11, no. 5, pp. 580-590, 2020.

[10] P. Diogo, N. V. Lopes and L. P. Reis, "An ideal IoT solution for real-time web monitoring", Cluster Computing, vol. 20, no. 3, pp. 2193-2209, 2017.

[11] M. M. Soto-Cordova, M. Medina-De-La-Cruz and A. Mujaico-Mariano, "An IoT based Urban Areas Air Quality Monitoring Prototype", International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, pp. 711-716, 2020.

[12] F. Kazemian and T. Howles, "Teaching Challenges - Testing and Debugging Skills for Novice Programmers", Software Quality Professional, vol. 11, no. 1, 2008.

[13] R. Chmiel and M.C. Loui, "Debugging: From Novice to Expert", ACM SIGCSE Bulletin, vol. 36, no. 1, 2004.

[14] J. Dolinay, P. Dostálek and V. Vašek, "Arduino Debugger", IEEE Embedded Systems Letters, vol. 8, no. 4, pp. 85-88, 2016.

[15] H. Li, Y. Xu, F. Wu and C. Yin, "Research of "Stub" remote debugging technique", Proceedings of 2009 4th International Conference on Computer Science & Education, Nanning, China, pp. 990-993, 2009.

Transliterating Nôm Scripts into Vietnamese National Scripts using Statistical Machine Translation

Dien Dinh[1], Phuong Nguyen[2], Long H. B. Nguyen[*3]
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

***Abstract*—Nôm scripts were used as the Vietnamese writing system from the 10th century to the early 20th century. During this period, Nôm scripts were the means to record a broad range of historical events, literary works, medical knowledge, as well as wisdom of many other domains. Unfortunately, since hardly any native Vietnamese speaker can read Nôm scripts nowadays, these valuable documents have not been fully harnessed. To address this gap, it is necessary to build an automatic transliteration system that can support us in decoding the ancient scripts and gaining knowledge of our Vietnamese ancestors. This study focuses on categorizing and reviewing the current progress on the Statistical Machine Translation (SMT) approaches to transliterate Nôm scripts into Vietnamese national scripts. In this paper, we discuss the differences between Nôm scripts and Vietnamese national scripts, systematically compare SMT models in transliterating Nôm scripts into Vietnamese national scripts, as well as having a thorough outlook on several promising research directions.**

***Keywords*—Statistical machine translation; automatic transliteration; Nôm Script (chữ Nôm); vietnamese national script (chữ Quốc ngữ)**

I. Introduction

Transliteration is a type of conversion of a text from one script to another, in the same language. For instance, the Cyrillic scripts of the Russian language, "Путин", is transliterated into the Latin scripts as "Putin". This transliteration is relatively straightforward, because there is only one correspondence in the Latin scripts for most of the letters in the Cyrillic scripts. Since both scripts are based on alphabets that contain a limited number of graphemes (strokes) to represent speech, transliteration can be done by looking up the mapping table. Table I is a portion of the mapping table from Cyrillic scripts to Latin scripts.

Table I. A Portion of the Mapping Table from Cyrillic Scripts to Latin Scripts

Cyrillic letter	Latin letter
А а	A a
Б б	B b
И и	I i
Н н	N n
П п	P p
Т т	T t
У у	U u

On the contrary, transliteration from Nôm scripts to Vietnamese national scripts is challenging because they do not belong to the same writing system. While Nôm scripts belong to the logographic writing system, Vietnamese national scripts belong to the alphabetic writing system. In other words, Nôm - Vietnamese national scripts is the one-to-many relationship. For instance, the Nôm character 𠊛 can be transliterated into *nghĩ* or *nghỉ*. Due to differences between the two writing systems, the mapping table method presented in the aforementioned Russian language example is not applicable when transliterating from Nôm scripts into Vietnamese national scripts.

The one-to-many mapping from Nôm scripts to the Vietnamese national scripts causes difficulties in transliterating process because people have to simultaneously read Nôm text and guess the appropriate meaning. Successful Nôm-transliteration also requires extra-linguistic knowledge about the culture, history, geography, dialects, specialized terminologies of ancient Vietnam. In recent years, rich-resource languages have gained success in applying machine translation. Chinese [1] and many European languages, including German [2], Greek [3], English [4], Spanish [5], French [6], Finnish [7], Italian [8], Dutch [9], and Portuguese [10] are some of those rich-resource languages. Besides, research in low-resource Southeast Asian languages such as Indonesian [11], Khmer [12], Lao [13], Malay [14], Myanmar [15], Philippines [16], and Thai [17], also yields significant results, which motivates us to apply machine translation in transliterating Nôm scripts into Vietnamese national scripts. Two state-of-the-art approaches in machine translation are Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). However, NMT requires a large amount of data [18], which is impractical for the low-resource language pair Nôm - Vietnamese national scripts. Therefore, we apply SMT for the transliterating task in this study.

Given the mechanism, the larger the manually transliterated training data are given to the computer, the more accurate the transliteration the computers generate. Besides, the machine can also improve the transliteration accuracy if humans supervise and manually revise the incorrect results that the computers previously produce. The more times we repeat the supervising and revising loop, the better the transliteration results become.

In this paper, we present the automatic transliteration from Nôm scripts to Vietnamese national scripts using Statistical Machine Translation. Our research steps are as the following: (1) collect and clean (i) the Nôm-Vietnamese national parallel corpus as the training data for the translation model and (ii) the monolingual Vietnamese national scripts as the training data for the language model, (2) classify corpora according to literary forms and domains, (3) experiment, and (4) analyze the experimental results.

Our main contributions are:

- Providing detail background about Nôm scripts

- Systematically comparing Nôm scripts with Chinese scripts
- Experimenting to show significance of the translation models in transliterating Nôm scripts into Vietnamese national scripts

The remaining of the paper is organized as follows: in Sections II and III, we provide an overview of Nôm scripts and of related studies, respectively. Then, we present our proposed model in Section IV and discuss the experimental results in Section V. Section VI concludes the study.

II. Overview of Nôm Scripts

Nôm scripts were created based on Chinese characters, which results in various similarities between Nôm scripts and Chinese scripts. Different from all phonological recording systems, Chinese scripts are the only logographical writing system currently used in the world [19]. Regarding the phonological writing system, there are symbols that record phonemes of a language. Meanwhile, Chinese characters are used to mark morphemes, ideas, basic concepts such as the sun (日), moon (月), tree (木), human (人), water (水), and heart (心). These basic elements are called radicals (部首). Radicals are the building blocks from which Chinese characters (Hanzi - 汉字) are built. According to the Han dictionary Shuowen (说文), there are six methods (六书) of constructing Chinese characters, including:

- Pictograms (象形): 日 (sun), 月 (moon), 木 (tree), etc.
- Ideogram (指事): 上 (above), 一 (one), 本 (root), etc.
- Combined ideogram (会意): 信 (trust), 林 (woods), 森 (forest), etc.
- Ideogram plus phonetic (形声): 妈 (mother) with 马(/mǎ/) as phonetic element and 女(/nǚ/) as ideographic element, etc.
- Derivative cognates (转注): 少 (a few - thiểu/to lack - thiếu), etc.
- Rebus (假借): 自 (oneself) which loans from the character 鼻(nose), etc.

Among these six types of characters, 90 percent belong to the ideogram-plus-phonetic category [19], i.e., each character is a *morpheme-syllable* compound. Meanwhile, Vietnamese language is constituted by *morpho-syllables*, which means *units* that constitute the two writing systems are equivalent. In the Chinese language, morphemes are radicals because a radical is the smallest meaningful unit of the Chinese writing system. Radicals are also the basis to arrange entries in Chinese dictionaries. For instance, to look up for the Chinese character 妈 (mother), we first search for the radical 女 (woman), since the character 妈 contains the radical 女. Then, we look up the remaining component, 马, by the number of strokes, which is three.

According to [20], while there are about 10,000 distinct pure morpho-syllables (not including transliterated morpho-syllables of loan words or scripts of ethnic languages) in Vietnamese, there are approximately 13,000 distinct Chinese characters (not including ancient characters, characters used for translitering loan words) in Chinese. Also from [20], each Chinese character has its own Unicode; the Chinese Unicode Charset is constructed based on various Chinese encoding charsets such as Big5 and GB; these encoding systems are gathered and aggregated into Unicode CJK charset; the first version of CJK was released in 1980 with roughly 13,000 Chinese characters; the number of encoded Chinese characters has grown over the years and reached 80,000 in 2018.

Most of the Nôm scripts were also created in the form of semantic (meaning)-phonetic (sound) compounds. The ancient Vietnamese usually borrowed two elements - one element for meaning and the other for the sound - from Chinese character collection to construct a Nôm character. For instance, the Nôm character 𠄩 means *number three*. In the Nôm character 𠄩, the Chinese character 巴, which has pinyin /bā/, denotes the sound, while the Chinese character 三 indicates the meaning. Similarly, in the Nôm character 爸, which means *father*, the Chinese character 巴 signifies the sound, while the Chinese character 父 expresses the meaning. Apart from the aforementioned semantic-phonetic compounds, there are a number of Nôm characters created by other methods, such as rebus, repetition, transfer, and diacritics adding. These methods signify the phonological difference between Nôm and Chinese characters [21].

Because Nôm scripts are mainly built on the semantic-phonetic compound method, there are cases in which one Nôm character is mapped to two or more Vietnamese national scripts. This typically happens when the national scripts have similar pronunciation and indicate synonymous meanings. This phenomenon can be explained by linguistic characteristics. While Vietnamese and Chinese languages are both tonal languages, they do not have the same number of tones. In particular, while there are six tones corresponding to six diacritics in Vietnamese, there are only four tones in Chinese. Moreover, different script creation methods, regional dialects, and Sino-Vietnamese variants due to different times of adoption also account for the one-to-many mapping between Nôm scripts and Vietnamese national scripts. For instance, the Nôm character 味 has two corresponding national scripts. The first one corresponds to *mùi* (smell), as it was adopted before the Tang Dynasty. Meanwhile, the second one corresponds to *vị* (flavor) as it was adopted from the Tang Dynasty onwards [22]. In the Nôm-Vietnamese national scripts dictionary[1], a considerable number of Nôm characters are **polyphonic** (a *polyphonic* Nôm character has more than one corresponding Vietnamese national script). For example, character 折 (Unicode code 6298h) has 19 corresponding national scripts (chệch, chét, chết, chẹt, chiết, chít, chịt, díp, gãy, gẩy, giẹp, giết, giỡn, nhét, nhít, siết, trét, triếp, xiết). This is also the one with the highest number of meanings in the Nôm-Vietnamese national scripts dictionary. In contrast, each **monophonic** Nôm character has only one corresponding Vietnamese national script. Table II are examples of polyphonic Nôm characters.

From the Nôm-Vietnamese national scripts dictionary, which includes 22,264 entries, we can classify Nôm characters according to the number of Vietnamese national scripts of each Nôm character, details are in Table III. According to the first row of the Table III, monophonic Nôm characters account for 76.654 percent of all dictionary entries. So, the remaining

[1]Hanosoft3. [Online]. Available: https://hanosoft-3-0-hanokey-2010.soft112.com. Accessed Jan 2019.

TABLE II. EXAMPLES OF POLYPHONIC NÔM CHARACTERS

Nôm script	Vietnamese national scripts	Quantity of Vietnamese national scripts
一	nhất, nhứt	2
丁	đinh, đĩnh	2
丐	cái, gái	2
万	muôn, vàn, vạn	3
与	dữ, dự, đử	3
丑	sấu, xấu, sửu	3
且	thả, vã, vả	3
世	thá, thé, thế, thể	4
中	đúng, trong, trung, trúng, truồng	5
丞	bậy, chăng, chẳng, phi, phỉ, vầy, vậy	7

23.346 percent are polyphonic Nôm characters. The proportion of polyphonic Nôm characters in comparison with monophonic Nôm characters is presented in Fig. 1.

TABLE III. FREQUENCY OF NÔM CHARACTERS ACCORDING TO THE NUMBER OF THEIR CORRESPONDING VIETNAMESE NATIONAL SCRIPTS

Quantity of corresponding national scripts	Quantity of Nôm character (Frequency)	Proportion
1	11,610	76.654%
2	1,907	12.591%
3	787	5.196%
4	384	2.535%
5	209	1.380%
6	94	0.621%
7	66	0.436%
8	30	0.198%
9	20	0.132%
10	16	0.106%
11	11	0.073%
12	3	0.020%
13	3	0.020%
14	2	0.013%
15	1	0.007%
16	1	0.007%
18	1	0.007%
19	1	0.007%
Total	**15146**	**100%**

Choosing the suitable Vietnamese national script for a given Nôm character is a difficult problem not only for the machine but also for the transliterators. Consider the Nôm character 𠊛 (Unicode code 2025Dh), which appears in the 12th sentence of Tale of Kieu in Fig. 2. 𠊛 might be transliterated into two national scripts as *nghĩ* (to think) or *nghỉ* (a pronoun used to indicate an old man in ancient Vietnamese) [23]. Scholars have been debating for over 50 years on which national script is correct in the given situation. Both sides provide various arguments, historical evidence, and literary evidence, etc. to demonstrate why one out of the two national scripts would be more suitable than the other. Therefore, requiring a computer to generate a 100-percent accurate transliteration output is impracticable, at least at present time and in near future.

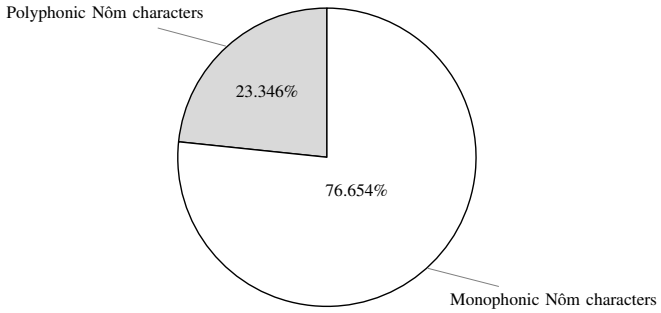


Figure 1. Proportion Polyphonic Nôm Characters versus Monophonic Nôm Characters (based on Table III).

家	資	𠊛	拱	常	常	堛	中
Gia	tư	**nghĩ /nghỉ**	cũng	thường	thường	bậc	trung

Figure 2. The 12th Sentence in Tale of Kieu by Nguyen Du.

III. RELATED WORKS

The digitization of Nôm scripts has been proposed and implemented since the 1990s by Ngo Thanh Nhan, Nguyen Quang Hong, among other scholars[2]. Thanks to these contributors, most of the common Nôm characters have become a part of the Unicode encoding system. This significant work is a solid foundation for lateral digitizing steps, such as storage, lookup, processing, automatic transliteration, etc.

Moreover, *Việt Hán Nôm 2002*, a software developed by Phan Anh Dung[3], allows us to type and look up both Chinese and Nôm characters. Another software, Hanosoft, developed by Tong Phuoc Khai[4], also includes several utilities for looking up and transliterating from Chinese characters into Nôm characters. In the aforementioned software, the authors have developed a tool to automatically transcribe Chinese characters into Sino-Vietnamese, Chinese characters into pinyin, and Nôm characters into national scripts. However, the central issue of the problem, which is choosing the proper National script for a given polyphonic Nôm character, has not yet been addressed. The software just randomly selects a Sino-Vietnamese phonetic transcript or a phonetic transcript among all possibilities. Besides, the website of the Vietnamese Nôm Preservation Foundation[5] includes a Chinese character-Nôm lookup tool and a digital library of Nôm documents, most of

[2]http://dir.vietnam.online.fr/home/vnChuNom.htm. Accessed May 2014.

[3]http://www.hannom.org.vn/detail.asp?param=507&Catid=363. Accessed Jun 2006.

[4]https://hanosoft-3-0-hanokey-2010.soft112.com. Accessed Jan 2019.

[5]http://www.nomfoundation.org. Accessed Oct 2019.

which are images of hand-written Nôm. Some literary works have also been digitized.

The work that is most closely related to our study is the Nôm converter[6], which is a toolkit used to automatically transliterate Nôm scripts to national scripts and vice versa. The system applies Statistical Machine Translation (SMT) approach and is based on Moses [24]. The data sets used to train Moses are parallel corpora. These corpora are 22 manually transliterated texts corresponding to 3,234 lines in total. The tool works fine, except for some cases in which input contains strange untrained Nôm scripts. For those cases, Nôm converter just ignores the strange untrained scripts and transliterates the rest of the input scripts as normal. Nôm converter has a rather high rate of choosing the correct national script when compared with the referenced transliteration version carried out by humans. To the best of our knowledge, it may be considered as the first automatic Nôm-Vietnamese national script transliteration tool that utilizes machine learning technology. Our approach is similar to Nôm converter, but with new modifications and improvements to address the limitations of the existing system.

IV. Proposed Model

In our proposed model, we customized a Statistical Machine Translation model (SMT) and improved the transliteration accuracy based on our work in automatic translation from English into Vietnamese [25]. Instead of following the Nôm converter system's approach in transliterating both directions (from Nôm scripts into Vietnamese national scripts and vice versa), we only focused on one-way transliteration from Nôm scripts into Vietnamese national scripts. Our core aim is to harness the Vietnamese ancient Nôm text, and the transliteration from national scripts to Nôm scripts does not imply as much practical significance. Besides, focusing on a one-way transliteration from Nôm scripts into national scripts allows us to invest more in improving national script output through various language models.

To overcome the shortage of parallel corpora for training as in the Nôm converter system, we added a Sino-Vietnamese dictionary into the phrase table of the Moses system. To improve the accuracy, we also added more manually transliterated literary works that Nôm converter has not yet included. Our major contribution is categorizing the Nôm script input data and providing language models for the Vietnamese national script output. The most challenging issue that we observed in transliterating Nôm script into Vietnamese national script was choosing the correct national script among all possibilities. This selection depends on context, form, domain, and even on the chronology of the input data. Nôm converter merely selects the national scripts according to the context in the training dataset, which is mixed in terms of form, domain, and time. Therefore, we classified the training dataset and language models by form and domain in our proposed model.

Because each form has its own rules for choosing the national script output, we classified the form into two categories: verse (such as Tale of Kieu, Tran Te Xuong's poems, etc.) and prose (The legend of Quynh, Biography of Phan Boi Chau, etc.). That is, these two forms required different language models. Besides, we also built corpora for three different domains, which were literature, history, and religion. New domains will be added into the current list of domains if we constructed and developed more corpora. Since each domain has its own terminologies, determining the domain to which the input scripts belong helped us narrow down the domain of possible national script output to improve the possibility of selecting the correct national scripts, especially for the cases in which the input is polyphonic Nôm scripts.

The final step was to build language models in the target language which was Vietnamese national scripts. The principle of machine learning is that the more training data we feed into the model, the better the transliterating accuracy will become. Due to this reason, not only did we utilize the national script dataset available in the parallel corpora that were used to train Moses in the previous step, but we also provided additional national script data that were already categorized by form and domain. This step improved the accuracy of the proposed model significantly since we included hundreds of thousands of sentences to build language models, compared to only thousands of sentences in the parallel corpora. A larger dataset of N-gram language models also allows the machine to generate the most linguistically natural transliteration output.

Later, when our proposed model is put into use, users will be able to select the form and domain of input data they want to transliterate from a menu. According to users' selection, computers will use the corresponding knowledge they have been trained to fit the form and domain of input data.

Let n be the source language sentence (Nôm scripts) and q be the target language sentence (Vietnamese national scripts), we have the following equation of the SMT model:

$$\hat{q} = \underset{q}{\operatorname{argmax}}\, P(q)P(n|q) \tag{1}$$

Through equation (1), the SMT model's working process is as follows:
i) estimate probability of seeing target string q language models $P(q)$;
ii) estimate probability that the source string n is the translation of the target string q given the translation model $P(n|q)$;
iii) choose the sentence q so that the value of product $P(q)P(n|q)$ is the maximum.
Fig. 3 shows a complete statistical translation system.

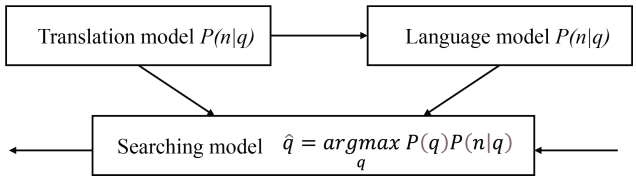


Figure 3. Transliteration Model

V. Experimental results and development directions

In this section, we present experimental results on our proposed model and compare the transliteration output with the baseline system Nôm converter. Then, we show limitations of

[6]Nôm converter. [Online]. Available: https://chunom.org/pages/moses/. Accessed Oct 2019.

our proposed model and corresponding development directions to improve those limitations.

A. Experiments and Results

In this sub-section, we describe the training data and experimental results of the proposed model.

1) Training, testing, and tuning datasets: We used single-character dictionaries listed in Table IV and compound-character dictionaries listed in Table V. In single-character dictionaries, each entry is one morpho-syllable such as 乂 - một (one), 仡 - là (to be). Meanwhile, entries in compound-character dictionaries have at least two morpho-syllables such as 乂匹 - một vài (several), 祝㦖辭潤 - chúc mừng năm mới (happy new year).

TABLE IV. SINGLE-CHARACTER DICTIONARIES USED FOR TRAINING

ID	Description	Size (Entries)	Source
1	Nôm - National Script dictionary	22264	Hanosoft3
2	Sino - Vietnames dictionary	16402	Hanosoft3
3	Sino (simplified) - Vietnames dictionary	10758	Hanosoft3
4	Sino (traditional) - Vietnames dictionary	11285	Hanosoft3
5	Nôm - National Script dictionary	32838	www.hannom-rcv.org
6	Nôm - National Script dictionary	879	www.chunom.org
7	Tale of Kieu (1902) dictionary	3353	www.nomfoundation.org
Total		97779	
After removing duplicates		**38897**	

TABLE V. COMPOUND DICTIONARIES USED FOR TRAINING

ID	Description	Size (Entries)	Source
1	Nôm - National Script dictionary	1951	www.chunom.org
2	Nôm - National Script dictionary	4520	www.hannom-rcv.org
Total		6471	
After removing duplicates		**6205**	

The Tale of Kieu (1902) dictionary in Table IV is not a publicly available dictionary. We observed there are Nôm characters in Tale of Kieu that have not been included in the other six single-character dictionaries. Therefore, we manually created the Tale of Kieu (1902) dictionary by listing all distinct pairs of Nôm and Vietnamese national scripts. We then utilized a computer program to aggregate all dictionaries listed in Table IV and Table V and removed the duplicate entries afterwards.

Regarding parallel sentences in the training and testing datasets, we used the corpus documents available on the websites chunom.org[7], Vietnamese Nôm Preservation Foundation[8], Việt Hán Nôm[9], and han-nom.org[10]. Details about domain and literary form of those sentences are listed in Table VI.

TABLE VI. NÔM-NATIONAL SCRIPT PARALLEL TEXTS IN TRAINING, TESTING, AND TUNING DATASETS

ID	Domain	Form	Size (Sentences)
1	Literature	Verse	7232
2	Literature	Prose	521
3	Religion	Verse	46
4	History	Verse	121
Total			7920

Currently, we have not utilized the domain information yet because the majority of the dataset belongs to the Literature domain. Classification of data is only useful if we have a considerably large corpus of various domains. Although we are not using categorizing information at the moment, we still include it into the program as a foundation for future work. We also collected corpora written in Vietnamese national scripts on the websites Gác Sách [11], Sách Phật giáo [12], and Ô Cửa Sổ [13]. Domain and form of monolingual corpora are listed in Table VII.

TABLE VII. MONOLINGUAL CORPORA USED TO TRAIN LANGUAGE MODEL

ID	Domain	Form	Size (Sentences)
1	History	Prose	257269
2	Religion	Prose	83535
3	Literature	Verse	29383
Total			370187

The training, testing, and tuning datasets are splitted by the ratio 1:1:8 as follows: for each text in Table VI, roughly 1/10 is distributed to the testing set, 1/10 is distributed in the tuning set, the remaining 8/10 is for the training set.

2) Experimental results: Using Moses SMT system [24], we conducted experiments on the corpora previously discussed and yielded the results in the Table VIII. From henceforth, Experiment 1 is abbreviated as Exp1, and Experiment 2 as Exp2, etc.

In Exp1, we measured the impact of the parallel corpus and the monolingual language model corpus on transliteration results with a single-character dictionary. With only a

[7]https://www.chunom.org. Accessed Oct 2019.
[8]http://www.nomfoundation.org. Accessed Oct 2019.
[9]http://hannom.huecit.vn. Accessed Oct 2019.
[10]http://www.han-nom.org. Accessed Oct 2019.
[11]http://www.gacsach.com. Accessed Jan 2020.
[12]http://www.sachphatgiao.net. Accessed Jan 2020.
[13]http://www.ocuaso.com. Accessed Jan 2020.

single-character dictionary equipped, the transliteration system behaves like a human with dictionaries, looking up a Nôm character in the dictionary and writing down a corresponding Vietnamese national script of that Nôm character. No linguistic knowledge was used and barely any understanding of Nôm script was required. The difference lies in the time it takes to transliterate Nôm script to national script. Humans might take days or weeks or even months to manually look up all Nôm characters in 786 lines of Nôm-script. However, machines take less than an hour if Moses is run on a high-spec machine. Since no linguistic knowledge was applied to the transliteration, the BLEU score [26] was 13.32. The results were acceptable, given the ambiguous nature of Nôm script. Because the model could not determine the context surrounding the input Nôm scripts, it could not choose the correct national scripts to generate a fluent output.

In Exp2, the data was the same as in Exp1, but the model has been tuned for better transliteration quality. The resulting BLEU score was 14.56, which was slightly better than the previous experiment's result.

In Exp3, we measured the impact of the language model by adding 370,817 lines of Vietnamese national script to train the model instead of using the default national scripts extracted from the parallel corpus. This made a significant difference, as the BLEU score increased from 13.32 to 63.89. This was because the language model supported phrase-based translation and provided context for the transliteration model to choose the most likely national script for a given Nôm script.

In Exp4, we tuned the model from Exp3, and the BLEU score increased from 63.89 to 65.94.

In Exp5, we added 6205 entries of compound dictionaries, growing the parallel corpus compared to that of Exp1. Consequently, the BLEU score increased from 13.32 to 36.82.

In Exp6, we tuned the model from Exp5, and the BLEU score increased from 36.82 to 44.24.

In Exp7, we added 370,817 lines of Vietnamese national script to train language model. Compared to the results in Exp5, the BLEU score in this experiment increased from 36.82 to 67.19.

In Exp8, we tuned the model from Exp7, and the BLEU score increased from 67.19 to 69.16.

In Exp9, we added 6,348 pairs to the parallel corpus. Compared to Exp1 and Exp5, the BLEU score increases from 13.32 to 36.82 to 80.50.

In Exp10, we tuned the model from Exp9, and the BLEU score increased from 80.50 to 80.83, which was not a considerable difference.

In Exp11, we added 370,817 lines of Vietnamese national script to train language models. Compared to Exp9, the BLEU score increased from 80.50 to 82.30. In this case, since we already had parallel corpus with long sentences of national script, adding a language model corpus did not yield a significant difference as in Exp3 and Exp7.

In Exp12, we tuned the model from Exp11, and the BLEU score increased from 82.30 to 85.38.

In the last four experiments, from Exp13 to Exp16, we used the parallel corpus without dictionaries to train the model and got acceptable results. However, there was similarity between the training corpus and the testing corpus, so the BLEU score was quite high, ranging from 75.71 to 79.40.

We hypothesized that if the testing data contain Nôm scripts from other domains such as medicine or agriculture, which have not been in the training data set yet, then the models with dictionaries will work better. This was based on the assumption that even though lacking the context, dictionaries cover a broader scope of vocabularies. However, that missing context could be made up by the additional language model as in Exp3 and Exp4, where the BLEU scores were 63.89 and 65.94 respectively. Those results were acceptable given that we trained the model only with dictionaries and additional language model data, without any parallel sentences. At the moment, we did not have data to verify our hypothesis. Verifying this hypothesis will be put in our future work, when we collect data from various other domains.

The corpora we used to train and test the transliteration system included single-character dictionaries, compound-character dictionaries, and parallel pairs of Nôm-Vietnamese national script sentences, whose model was trained by dictionaries. Parallel sentences were separated with the ratio of train:tune:test as 8:1:1 (tune here refers to the data used to tune the model, that is, to find the optimal parameters for the transliteration model).

In the third column of Table VIII, "Default" refers to the monolingual national scripts extracted from the parallel corpus in the second column. In experiments with "Default" monolingual corpus, we did not use additional language model corpora in Table VII. In the fourth column of Table VIII, BLEU stands for Bi-Lingual Evaluation Understudy, a metric used to measure quality of machine translation output in comparison to human-generated output. Format of BLEU score is **overall**, uni-gram/2-gram/3-gram/4-gram. The fifth column signifies whether an experiment was tuned or not. As mentioned previously, the purpose of tuning is to find optimal parameters for the transliteration model, and thereby generating better transliteration output, compared to the untuned model.

After training the model, we chose 10 percent of the sentences in the testing data to evaluate the proposed model. Exp12 yielded the highest BLEU score, which was 85.38. Therefore, we selected some sentences in the testing set of this experiment to compare to the corresponding output generated by Nôm converter. 12 sentences from Tale of Kieu (version 1902) were tested, and the results are presented in Table IX.

We use different typefaces to distinguish between correct and incorrect transliteration. The differences are explained as follows:

- **Compared transliteration**
- Correct transliteration
- *Synonymical transliteration*
- ***Incorrect/un-handled transliteration***

TABLE VIII. EXPERIMENT RESULTS

Exp ID	Training Data		BLEU	Tuned
	Parallel Corpus	Monolingual Corpus		
1	Table IV	Default	**13.32**, 44.6/19.6/8.9/4.0	No
2	Table IV	Default	**14.56**, 47.1/22.1/9.6/4.5	Yes
3	Table IV	Table VII	**63.89**, 84.4/69.3/57.6/49.6	No
4	Table IV	Table VII	**65.94**, 83.4/71.3/60.6/52.5	Yes
5	Table IV, Table V	Default	**36.82**, 67.5/45.6/29.7/20.1	No
6	Table IV, Table V	Default	**44.24**, 71.9/52.1/37.5/27.3	Yes
7	Table IV, Table V	Table VII	**67.19**, 85.0/72.3/61.4/54.0	No
8	Table IV, Table V	Table VII	**69.16**, 85.3/74.2/64.1/56.5	Yes
9	Table IV, Table V, 6348 sentence-pairs	Default	**80.50**, 91.3/84.1/76.7/71.4	No
10	Table IV, Table V, 6348 sentence-pairs	Default	**80.83**, 91.5/84.5/77.2/71.6	Yes
11	Table IV, Table V, 6348 sentence-pairs	Table VII	**82.30**, 92.2/85.4/79.1/73.7	No
12	Table IV, Table V, 6348 sentence-pairs	Table VII	**85.38**, 93.3/88.0/82.8/78.3	Yes
12.2.1	Table IV, Table V, 6348 sentence-pairs	Table VII, 6348 sentences	**85.76**, 93.9/88.5/83.1/78.5	No
12.2.2	Table IV, Table V, 6348 sentence-pairs	Table VII, 6348 sentences	**85.71**, 93.6/88.4/83.1/78.6	Yes
13	6348 sentence-pairs	Default	**77.04**, 89.3/81.0/73.0/66.9	No
14	6348 sentence-pairs	Default	**77.01**, 89.2/80.9/73.0/66.8	Yes
15	6348 sentence-pairs	Table VII	**75.71**, 88.7/79.7/71.6/65.0	No
16	6348 sentence-pairs	Table VII	**79.40**, 90.2/82.9/75.9/70.2	Yes

TABLE IX. TRANSLITERATION OUTPUT OF PROPOSED MODEL IN COMPARISON WITH NÔM CONVERTER

Nôm input sentences	Referenced transliteration	Proposed model	Nôm converter
聂辭融堺趴些	trăm năm trong cõi người ta	trăm năm trong cõi người ta	trăm năm trong cõi người ta
笋才笋命窖罕恄儭	**chữ** tài **chữ mệnh** khéo **là** ghét nhau	chữ tài chữ mệnh khéo là ghét nhau	*笋* tài *笋 mạng* khéo *呀* ghét nhau
黫戈沒局滌桄	**trải** qua **một** cuộc bể **dâu**	trải qua một cuộc bể dâu	*黫* qua *沒* cuộc bể *桄*
仍調韹凳缶疥疽恚	những điều trông thấy đã đau **đớn** lòng	những điều trông thấy đã đau đớn lòng	những điều trông thấy đã đau *疽* lòng
迍咦彼酋斯豐	**lạ** gì **bỉ sắc tư phong**	lạ gì bỉ sắc tư phong	*呀* gì *彼酋 斯 豐*
荃撐悁貝膦紅打慳	trời xanh quen với má hồng đánh ghen	trời xanh quen với má hồng đánh ghen	trời xanh quen với má hồng đánh ghen
稿贅吝捵趧畑	**cảo** thơm lần **giở** trước đèn	cảo thơm lần giở trước đèn	*稿* thơm lần *捵* trước đèn
風情固錄羣傳史撐	phong tình có **lục** còn **truyền** sử xanh	phong tình có lục còn truyền sử xanh	phong tình có *錄* còn ***truyện*** sử xanh
浪辭嘉靖朝明	rằng năm gia **tĩnh triều** minh	rằng năm gia tĩnh triều minh	rằng năm gia *靖* ***chiều*** minh
罘方滂瀙缶京凭鑌	bốn phương phẳng **lặng** hai kinh vững vàng	bốn phương phẳng lặng hai kinh vững vàng	bốn phương phẳng ***láng*** hai kinh vững vàng
固茄員外尸王	có nhà viên ngoại họ vương	có nhà viên ngoại họ vương	có nhà viên ngoại họ vương
家資俙拱常常堛中	gia tư **nghĩ** cũng thường thường bậc trung	gia tư *nghĩ* cũng thường thường bậc trung	gia tư nghĩ cũng thường thường bậc trung

The BLEU score in Exp12.2.1 was the highest score. However, the way we separated data into training set and testing set previously might cause biased results because of the similarity between the training set and testing set. That is, it may not be practical to distribute each poem (text) in both training and testing sets with the ratio of 8:1, because in real world situations, users might want to transliterate an unseen Nôm text, completely different from the one used to train our model. Therefore, we applied k-fold cross validation to better evaluate the skills of our proposed model. There are 7,920 lines of parallel Nôm-Vietnamese national scripts in total. We shuffled the data and then distributed parallel text into 10 folds (parts). After equally distributing all sentence-pairs into 10 folds, each fold contained 792 pairs of sentences. Based on the experiment results presented in Table VIII, we observed that Exp12.2.1 set-up generated the highest transliteration quality. Consequently, we implemented k-fold cross validation using dictionaries and an additional language model for model training as in Exp12.2.1. Then 10 previously separated folds were

distributed into training, tuning, and testing sets as follows: eight folds for training the model, one fold for tuning, and the one remaining for testing. This time, the data used for language model training was slightly different from that of Exp12.2.1. In addition to the data as in Exp12.2.1, we also extracted and used the national scripts from eight folds of the parallel corpus to feed more data into the model, and thereby improving the transliteration quality generated from the model. We conducted the experiment 10 times, with the corresponding BLEU evaluations presented in Table X. Averaging BLEU

TABLE X. EXPERIMENTS AND RESULTS OF K-FOLD CROSS VALIDATION

Exp ID	BLEU	Exp ID	BLEU
1	**81.88**, 92.0/85.2/78.8/73.1	6	**83.32**, 92.6/86.5/80.3/75.2
2	**83.47**, 92.5/86.5/80.7/75.3	7	**83.83**, 92.6/86.6/81.0/76.1
3	**83.85**, 92.6/86.7/81.0/76.2	8	**83.00**, 92.3/86.2/80.0/74.5
4	**83.65**, 92.7/86.6/80.7/75.6	9	**82.35**, 92.0/85.5/79.2/74.1
5	**83.11**, 92.4/86.4/80.2/74.7	10	**82.67**, 92.1/85.9/79.7/74.6
Average		**83.10**	

scores of 10 experiments, we got 83.10, which was quite close to our best result of 85.76 in Exp12.2.1, Table VIII. We conclude that the experiments carried out in Table VIII were relatively fair, as they were not biased due to the data distribution among the training set and the testing set.

B. Limitations and Development Directions

Based on the test results presented in S ection V-A, we observe that our proposed model still has limitations in choosing the correct national script for a given Nôm script input. While our goal to resolve this difficulty remains, it is unlikely to attain 100-percent accurate transliteration output since even humans argue over which national script should be used for a given Nôm character.

To overcome the aforementioned limitations, we will continue to collect and build a larger parallel corpus for the translation model as well as a monolingual corpus for language models. We will also categorize input data into domains to improve the transliteration quality. We will keep collecting corpora from some other domains such as medicine and ideology. In addition, we will conduct more experiments and train our proposed model with new machine learning models.

VI. CONCLUSION

In this paper, we have presented an automatic transliteration from Nôm scripts into Vietnamese national scripts using the SMT paradigm in computational linguistics. Our proposed model demonstrates significant improvements compared with the existing transliteration system, Nôm converter. Not only does the model recognize a broader range of Nôm scripts, it can also choose the national script for a given Nôm character with higher accuracy according to the context of the input Nôm scripts. Our finding of the distinct characteristic of the language pair Nôm - Vietnamese national scripts and our contribution in building a separate corpus for the language model beside the default language model extracted from the parallel corpus lead to a high result in the SMT approach.

In the future, we will build domain-specific language models and integrate linguistic knowledge to improve transliteration accuracy. Moreover, we can conduct manual post-editing to introduce further improvement. Our proposed model, therefore, will be able to generate more accurate transliteration results. This automatic transliteration system will bridge the gap between our past and our present, stemming the differences in our two writing systems, the historical Nôm scripts and our current national scripts. Thanks to this system, the priceless treasure of our ancestors in history, literature, religion, geography, and traditional medicine will be explored and harnessed effectively. Scholars can now browse and understand the main ideas of a Nôm text without having to invest an immense amount of time to manually work on the ancient scripts.

REFERENCES

[1] S. Liu, L. Wang, and C.-H. Liu, "Chinese-portuguese machine translation: A study on building parallel corpora from comparable texts," 04 2018.

[2] J. Marx, N. Smith, and Staudinger, "Some problems in the evaluation of the russian-german machine translation system miroslav," 02 2021.

[3] O. Nikolaenkova, "Applying clp to machine translation: A greek case study," *Journal of Applied Linguistics and Lexicography*, vol. 1, pp. 69–78, 09 2019.

[4] Y. Eytani, A. Lavie, E. Peterson, K. Probst, and S. Wintner, "Hebrew to english machine translation," 02 2021.

[5] M. Crespo and M. Sánchez-Saus Laserna, "Graded acceptance in corpus-based english-to-spanish machine translation evaluation," 01 2016.

[6] F. Bouzit and M. T. Laskri, "Arabic to french machine translation system based on dcf approach," 02 2021.

[7] G. Tang, R. Sennrich, and J. Nivre, "Understanding pure character-based neural machine translation: The case of translating finnish into english," 12 2020.

[8] R. Tse, S. Mirri, T. Su-Kit, G. Pau, and P. Salomoni, "Building an italian-chinese parallel corpus for machine translation from the web," 09 2020, pp. 265–268.

[9] R. Cornet, C. Hill, and N. de Keizer, "Comparison of three english-to-dutch machine translations of snomed ct procedures," *Studies in health technology and informatics*, vol. 245, pp. 848–852, 01 2017.

[10] Y. Li, C. Pun, and F. Wu, "Portuguese-chinese machine translation in macao," 02 2021.

[11] M. Dwiastuti, "English-Indonesian neural machine translation for spoken language domains," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 309–314. [Online]. Available: https://www.aclweb.org/anthology/P19-2043

[12] Y. Kyaw Thu, V. Chea, A. Finch, M. Utiyama, and E. Sumita, "A large-scale study of statistical machine translation methods for Khmer language," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, Oct. 2015, pp. 259–269. [Online]. Available: https://www.aclweb.org/anthology/Y15-1030

[13] A. Srithirath and P. Seresangtakul, "An approach to lao-english rule based machine translation," *Proceedings of the 2015-7th International Conference on Knowledge and Smart Technology, KST 2015*, pp. 93–98, 02 2015.

[14] S. Ab, N. Abdul Rahman, and N. Aziz, "Improving word alignment in an english – malay parallel corpus for machine translation," 02 2021.

[15] M. Zin, T. Racharak, and N. Le, "Construct-extract: An effective model for building bilingual corpus to improve english-myanmar machine translation," 01 2021, pp. 333–342.

[16] N. Oco and R. Roxas, "A survey of machine translation work in the Philippines: From 1998 to 2018," in *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*. Boston, MA: Association for Machine Translation in the Americas, Mar. 2018, pp. 30–36. [Online]. Available: https://www.aclweb.org/anthology/W18-2204

[17] S. Lyons, "A review of thai–english machine translation," *Machine Translation*, vol. 34, 09 2020.

[18] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. [Online]. Available: https://www.aclweb.org/anthology/W17-3204

[19] H. Rogers, *Writing Systems: A Linguistics Approach*, 1st ed. Blackwell, 2005.

[20] D. Dinh, *Từ điển học tính toán*. VNU-HCM, 2019.

[21] T.-C. Nguyen, *Diên cách cấu trúc chữ Nôm Việt*. VNU-Hanoi, 2012.

[22] K. D. Le, *Từ vựng gốc Hán trong tiếng Việt*. VNU-HCM, 2002.

[23] H. T. Le, "Nghĩ về một số từ khó hiểu trong Truyện Kiều," *Kiến thức ngày nay*, Jan 2016.

[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://www.aclweb.org/anthology/P07-2045

[25] D. Dinh, K. Hoang, and E. Hovy, *BTL: an Hybrid Model in the English – Vietnamese Machine Translation System*. Proceedings of the MT Summit IX, 2003.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040

Improve the Effectiveness of Image Retrieval by Combining the Optimal Distance and Linear Discriminant Analysis

Phuong Nguyen Thi Lan[1]
Thai Nguyen University – Lao Cai Campus
Lao Cai, Vietnam

Tao Ngo Quoc[2]
Institute of Information Technology
Vietnam Academy of Science and Technology
Hanoi, Viet Nam

Quynh Dao Thi Thuy[3]
Faculty of Information Technology
Posts and Telecommunications Institute of Technology
HaNoi, Viet Nam

Minh-Huong Ngo[4]
Institute of Sciences of Digital, Management and Cognition
University of Lorraine, France

***Abstract*—In image retrieval with relevant feedback, classification and distance calculation have a great influence on image retrieval accuracy. In this paper, we propose an image retrieval method, called ODLDA (Image Retrieval using the optimal distance and linear discriminant analysis). The proposed method can effectively exploit user's feedback from relevant and irrelevant image sets, which uses linear discriminant analysis to find a linear projection with an improved similarity measure. The experimental results performed on the two benchmark datasets have confirmed the superiority of the proposed method.**

***Keywords*—Content-based image retrieval; deep learning; similarity measures; Mahalanobis metric distance; linear discriminant analysis**

I. Introduction

Due to the need to efficiently process huge and rapidly increasing amounts of multimedia data, content-based image retrieval (CBIR) has received a lot of attention from researchers over the past few decades. Many CBIR systems have been developed, including QBIC [21], Photobook [22], MARS [23], PicHunter [24], Blobworld [25], SIMPLIcity [26].

In a typical CBIR system, low-level visual features include color, texture, and shape, which are automatically extracted and represented as feature vectors. It should also be added that feature vectors are good if they are of the high semantic meaning of the image and serve well for image comparison. To find the desired images, the user gives a sample image and the system returns a list of similar images based on the extracted features. When the system presents a list of images that are similar to the query image, the user marks the images most relevant to the given query image to get a feedback list. The system relies on this feedback list to learn a representation or similar measure to improve the accuracy of the image retrieval.

Therefore, the representation of the image by the feature vector and the similarity measure are the two main factors that influence the efficiency of the CBIR system. Improving the effectiveness of the CBIR system is a challenging issue in research. To improve efficiency, we need to reduce semantic gaps in CBIR. The semantic gap implies the difference between the image represented by the low-level feature that is automatically extracted and the semantics of the human perceived image. To reduce this semantic gap, we need to incorporate machine learning into the image retrieval process.

Recently, there are good results due to the application of CNNs to CBIR. It has been shown that if a CNN is trained in a full surveillance context on a large set of object recognition tasks, the features extracted from the CNN can address a variety of tasks such as object image classification, scene recognition, attribute detection, and image retrieval [27,28]. Research in [29] has shown that the performance of CBIR systems using CNNs is competitive even when CNNs are trained for an unrelated classification task. To improve efficiency right from the process of building an image representation feature set, the proposed method will use CNN to build a high semantic feature set. Besides, the proposed method will incorporate similarity metrics learning techniques to have an improved similarity measure more consistent with the data.

The idea of learning similarity metrics is to find an optimal distance measure that minimizes the distance between pairs of similar images and maximizes the distance between pairs of dissimilar images. This optimal distance measurement is then used to re-rank the entire set of images and return better results. In this paper, we propose an effective image retrieval technique, called ODLDA (Image Retrieval using the optimal distance and linear discriminant analysis). The proposed method is more accurate than some state of the art methods because the feature representation is highly semantic and the similarity metrics being learned is consistent with the data. By experimenting with two databases, we will show the accuracy of the proposed method.

The remainder of the paper is organized as follows. Section 2 reviews some related studies. We present in detail the proposed method in Section 3. Section 4 describes and

analyzes our experimental results. Setion 5 concludes this paper.

II. Related Work

Learning similar metrics in content-based image retrieval has received the attention of the research community [6,9,13,14,15,16,17,18]. In image retrieval with relevant feedback, the input data of distance learning algorithms are often divided into two groups: the first group consists of pairs of similar images; the second group consists of pairs of similar images and the pairs of images are not similar.

The idea of adjusting the weights of the distance function has been included in some content-based image retrieval methods such as SRIR [19]. These methods often take advantage of information from pairs of similar images and consider the scattering of the data on each dimension to construct an improved Euclidean distance function.

The MCML method [4] learns a Mahalanobis distance measure so that samples of the same class will be mapped to the same point. The distance metric learning problem is referred to as the convex optimization problem and is solved by the Gradient Descent method. However, the limitation of this method is the large computational complexity because it uses the Gradient Descent method to solve the convex optimization problem.

The idea of the LMNN [5] method is to minimize the distance of the samples of the same label in K-Nearest Neighbor and to maximize the distance of the samples that are not of the same label by a larger margin. It uses the Mahalanobis distance function. This idea is expressed as an optimization problem and solved by the SDP method [3] to find the improved distance metric.

Online Algorithm for Scalable Image Similarity learning (OASIS) [18] is specifically designed to work with pair constraints. However, they are based on strong assumptions about the input data or the structure of the constraints (requiring the input data to be sparse vectors). Therefore, it is difficult to apply in practice.

The idea of the Xing method [20] is to attribute to the convex optimization problem that minimizes the total distance of similar image pairs with the constraint that the total distance of pairs of images that are not similar reaches the maximum. In the initial phase, the method using the Euclidean distance function is improved with $A = I$. The Xing method presents an improved distance function where A is the result of the convex optimization problem. However, Xing's method has a large computational complexity due to the use of the Gradient Descent method and has not yet exploited information of similar image pairs.

The idea of the RCA method [8] is to use only similar pairs, find a data transformation based on a matrix of variance that is generated from pairs of similar images. From there it improved the Mahalanobis distance function by altering the weighting matrix. Although this method has lower computational complexity than that of the Xing method, however, the RCA method is limited to only considering the same set of images.

From analyzing the limitations of the above-related works, we propose an improved image retrieval method with an improved distance function. Improvement of the distance function which is based on maximizing the quotient between the total distance of dissimilar image pairs and the total distance of similar image pairs. Here, we look at both similar and dissimilar image sets to find the weight matrix and improve the efficiency of the retrieval method.

III. Proposed Image Retrieval Method

In this section, we will briefly present our proposed method. First, our proposed method builds deep features for representing images. Next, on the result set of the initial retrieval phase that uses deep features, the user marks up the images that are related to the query image to obtain the relevant image set (including relevant samples and samples are not irrelevant to the query image). Based on the relevant sample set, the proposed method is to train the model to find the linear projection. This linear projection satisfies the condition that the variance between samples in the same relevant set is minimized while maximizing the variance between the relevant and irrelevant samples. Besides, our proposed method also builds an improved Mahalanobis similarity metric by finding the optimal matrix M in the improved similarity metric formula.

A. Overview of the Proposed Method

A diagram of the proposed ODLDA, method is shown in Fig. 1. The method of using the CNN model has been trained on an ImageNet data set to extract the deep feature (high-level feature). When a user submits a query image, the method of extracting the deep feature of the query image is in the same way as performing an extraction with a database image. It then compares the similarity between the query image feature vector and the feature vector set of the image database which uses the Euclidean distance to return the initial result set to the user. Users conduct feedback by marking the images that are relevant and irrelevant to the query image to obtain the feedback image set. Then the feedback image set is used as input to the weight optimization and distance metric learning algorithm. Next, all images that are in the image database are re-ranked, which are based on the value of the improved Mahalanobis distance function. If the user is not satisfied with the result set, the feedback process will be repeated. If the user is satisfied, the system returns the final result set to the user.

B. Represent Image Features using Deep Learning

In recent years, CNN network has brought great results in the field of machine vision such as image classification problem, object identification, semantic segmentation. On that basis, there are many studies on content-based image retrieval using CNN and have obtained good results.

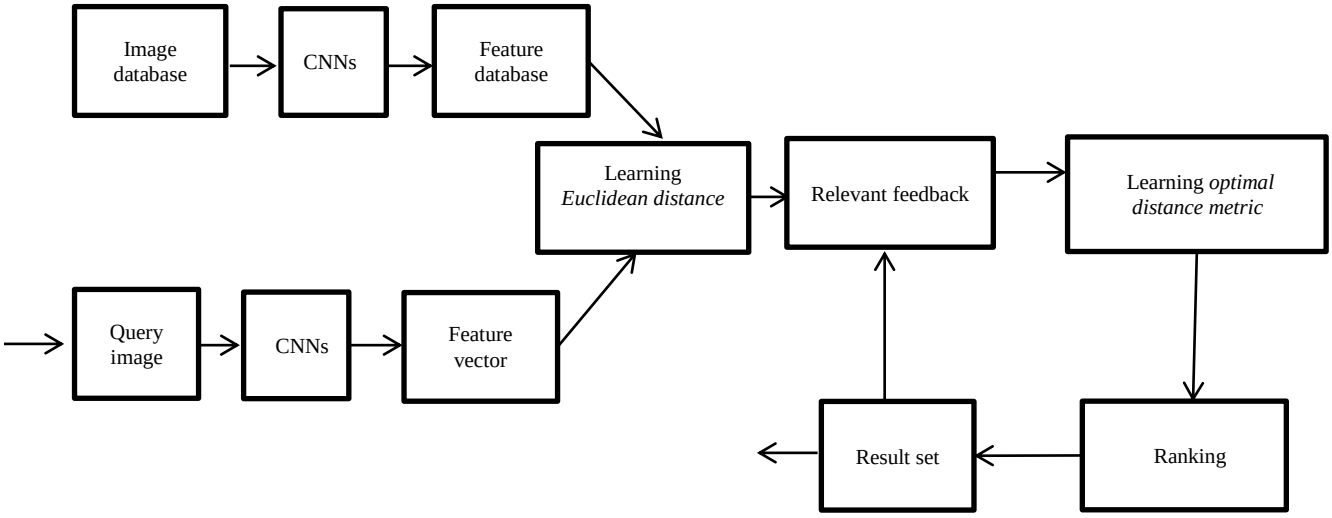


Fig. 1. Diagram of ODLDA Method.

In the document [1,2,7] has shown several approaches to improve the efficiency of a CBIR system using deep learning in building a more semantic feature set: 1) uses a pre-trained CNN model to construct an image feature set with an L_2 distance to compare the similarity measures between feature vectors; 2) it still uses the pre-trained CNN model to build the feature set, but improves it by using distance metric learning (DML) to obtain a similarity metric that is better suited to the data; 3) With a specific data set, retraining the CNN model associated with a specific classifier, then using the metric as 1) or 2) approaches is to complete a retrieval method.

Assuming we have two images in the database, I_i and I_j, the deep features are extracted using a pre-trained CNN model on the Imagenet dataset. The high-level feature of the two images I_i and I_j is denoted by x_i and x_j. The similarity metric used to compare these two features is L_2:

$$similarity(x_i, x_j) = \|x_i - x_j\|_2$$

$$= \sqrt{(x_i - x_j)^T (x_i - x_j)} \quad (1)$$

Formula (1) shows the similarity between images I_i and I_j, the greater the similarity, the more similar images I_i and I_j are.

Similarity metric using approach 2) to compare two feature vectors of the image calculated by the formula L_T:

$$similarity(x_i, x_j) = \|x_i - x_j\|_T$$

$$= \sqrt{(x_i - x_j)^T T (x_i - x_j)} \quad (2)$$

With a matrix, T obtained from learning the similarity metric which satisfies the condition T is a positive defined matrix, because the similarity metric must be positive, and the similarity metric has the smallest value when $x_i = x_j$.

The similarity metric here is that in approach 1) when the matrix T is a unit matrix $T = I$. In other words, it is a special case when we consider the correlation between the feature components in approach 1). Furthermore, each feature component has a different similarity, so it is often the similarity metric with approach 2) to get higher efficiency.

The proposed method is to build feature sets based on deep learning. After performing the K-NN procedure to obtain a list of initialization results and return them to the user, the user will mark the images that are related to the query image to obtain the feedback set. Next, it constructs an improved similarity metric by utilizing the positive sample set, which is inspired by approach 2) to construct the matrix T in the similarity metric formula (2). Matrix M is a complete matrix, which reflects the correlation of data on each feature and between features.

In the proposed method, we use a pre-trained CNN model on a very large data set. It then uses the model to extract high-level features, also known as image representation learning. The main reason we choose this approach is that a large enough data set is not available to train a CNN, Also, to train a CNN model, we will need a lot of time. CNNs are commonly used for image classification problems, in which an image is propagated across the network and the final probability is taken from the bottom layer of the network. However, in the process of learning a representation, instead of allowing the image to propagate over the entire network, we can stop the transmission at an arbitrary layer, for example, the final fully connected layer, and extracts the values from the network at this point, then uses them as feature vectors.

In the proposed method, we only use convolutional layers to extract features. The aim is to generalize a pre-trained CNN in learning the specific features of the image in the data set. The pre-trained model is used to obtain more powerful feature vectors than some algorithms such as SIFT, GIST, HOG, etc. We exploit the ability of a widely known convolutional neural network model, called ImageNet, pre-trained in ILSVRC 2012 with 1.2 million images and 1000 concepts to acquire outstanding features of the image. It consists of convolutional layers, pooling layers, and fully connected layers. The preceding layers are usually Convolutional layers combined with nonlinear activation functions and pooling layers (collectively referred to as ConvNet). The last layer is a fully-connected layer and is usually a softmax regression (see Fig. 2). The number of units in the last layer is equal to the number of layers (with ImageNet is 1000). So the output near

the last layer can be considered as a useful feature vector and Softmax Regression is the classifier used. The model uses a fixed size 256 x 256 input, while the data set used in the proposed method has a variable size of images. Therefore, the images are preprocessed by converting them to 256 x 256 size. When using the network to extract the fixed feature, we cut the network at a point before the last fully connected layer. Therefore, we obtained a feature vector of 1000 dimensions for each image.

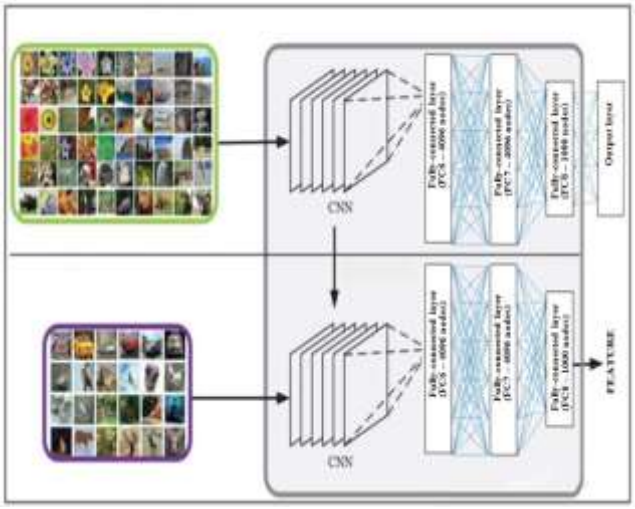


Fig. 2. Representational Learning Architecture is based on the Pre-Training of the CNN Model.

C. An Improved Distance Metric

Up to now, there have been several different distance learning methods that exploit the properties of the user feedback set during image retrieval. However, existing methods generally consider only the positive sample set but ignore the negative sample set. The basic idea of linear discriminant analysis (LCA) is to find an optimal transformation leading to an optimal distance function, which is accomplished by maximizing the sum of variance between samples of different classes (negative or positive) and minimize the variance of data in the same class (negative or positive).

Assume that the initial resulting set consists of N images: $X = \{x_i\}_{i=1}^{N}$. The initial result set is returned to the user's feedback and is divided into two distinct sets: a positive sample set and a negative sample set. To achieve the goal, we need to define two matrices of variance, S_b and S_w. Where, S_b is the distance between the expectations of the different classes and S_w is the distance between the expectations and the samples of each class. These two matrices are calculated by the formula:

$$S_b = \frac{1}{n_b}\sum_{j=1}^{2}\sum_{i\in D_j}(m_j - m_i)(m_j - m_i)^T \quad (3)$$

$$S_w = \frac{1}{n}\sum_{j=1}^{2}\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji} - m_i)(x_{ji} - m_i)^T \quad (4)$$

Where n_b is the total number of samples of the two sets of positive and negative samples, m_j is the center of class j, x_{ji} is the ith vector of class j, each D_j is a class. In this problem, we have 2 classes: positive class and negative class. Center m_j of class j is calculated by the formula: $m_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji}$.

The LDA process is referred to as the optimal problem as follows:

$$T = argmax_T \frac{|T^T S_b T|}{|T^T S_w T|} \quad (5)$$

Matrix T is the optimal transformation matrix, which we need to find. When we obtain the optimal transformation T, we get the optimal weight of the Mahalanobis distance function: $M_o = T^T$T.

According to the Fisher theory [11,12], the optimization problem (5) is equivalent to maximizing the total expected distance of different classes ($\hat{C}_b$) and minimizing the total expected distance in the same class (S_w) [10]. To find the solution to the problem (5), we propose to apply algorithm 1.1 below. This algorithm is also used to solve for previous studies on LDA [22].

D. Image Retrieval Algorithm

Algorithm 1.1, called ODLDA (Image Retrieval using the optimal distance and linear discriminant analysis) describes an effective image retrieval algorithm based on the optimal distance and linear discriminant analysis.

Algorithm 1.1. ODLDA

Input:

Image set: ***DB***

Initialization query image: **Q**

Returned image number for each iteration: ***N***

Output:

Result set: ***R***

1. S ←IRL<DB,*M*>;

2. S_q ←IRL<Q,*M*>;

3. $Result_{Initial}(Q)$←***Retrieval*** $_{Initial}(S_q, S, N)$

4. R←$Result_{Initial}(Q)$;

5. **Repeat**

5.1. $< F_{feature}, F^{+}_{label}, F^{-}_{label} >$)←***Feedback*** (R) ; *relevant feedback*

5.2. $A =$ ***LDA***$(F_{feature}, F^{+}_{label}, F^{-}_{label})$; *Find the optimal transformation* ***T***

5.3. $M_o = T^T T$; *The optimal weight of the Mahalanobis distance function*

5.4. *R*←***Ranking*** (S, M_o, N)*; Rerank the set of images according to the Mahalanobis distance function with the optimal weight*

until (User stops responding);

6. **Return** R;

The ODLDA algorithm is implemented as follows: Each image in the DB image set is represented by a feature vector in multidimensional feature space (Step 1). When the user introduces an image of the initialization query Q, the algorithm represents the query image into a feature vector S_q (Step 2). The initialization query is performed in Step 3 by $Result_{Initial}(Q) \leftarrow Retrieval_{Inital}\left(S_q, S, N\right)$, where S_q is the representation of the query image, S is the representation set of the database image set and N is the number of images to be retrieved in set S after each iteration. The retrieval result with the initialization query $Result_{Initial}(Q)$ is assigned to R (Step 4).

On the $Result_{Initial}(Q)$ set returned by the initialization query, the user responds through the function $Feedback(R)$ to get the feature set $F_{feature}$ and the label set $F_{Label} = \{F^{+}_{label}, F^{-}_{label}\}$ (Step 5.1). The user's feedback, including the relevant and irrelevant feedback set, is then fed into LDA (Step 5.2) to find projection A. Finding the projection A is done by solving the optimization problem (5). The results of this projection matrix were included to construct the optimal weight matrix to improve the weight of the Mahalanobis distance function (Step 5.3). At this point, we obtain the following improved Mahalanobis distance function:

$$d_M(F_i, F_j) = \|F_i - F_j\|_M = \sqrt{(F_i - F_j)^T M (F_i - F_j)}$$

The retrieval process reclassifies the entire image set in the image database by the function Ranking $(S, M,)$ N, and takes N images as the result set returned to the user (Step 5.4).

IV. Experimental Results

A. Experimental Environment

1) Image Dataset COREL: The image set that we used for our experiment is Corel Photo Gallery with 10800 images Fig. 3. Some of the topics for this set[1] include bonsai, castle, cloud, autumn, aviation, dog, primate, ship, stalactite, fire, tiger, elephant, iceberg, train, waterfall, Each image in this set contains a prominent foreground object. Each topic consists of about 100 images. The size of the images is 120 * 80 or 80 * 120.

2) Ground truth for evaluating the precision of the CBIR: Ground truth set is used to evaluate the precision of the CBIR system, i.e., the relevant or irrelevant images identified under this set. Accordingly, the image retrieval system considers the images that are related to the query image as images with the same subject. This set consists of 3 columns (titled: Query Image ID, Image ID, and Relation) and consists of 1,981,320 rows.

3) Image Dataset SIMPLIcity: To demonstrate the performance of the proposed method, in addition to experimenting on Image Dataset COREL, we also conducted experiments on Dataset SIMPLIcity. This is a small data set with a thousand images and 10 categories. Each image in this set is 256×384 or 384×256. Some samples in this image database are shown in Fig. 4. We represent each image by two features, that is, color and edge features. The color feature is represented by the color structure descriptors with a 128-dimensional vector, while the edge feature is the edge histogram descriptors with the 150-dimensional vector. A vector of 278 dimensions, composed of two color and edge features, represents an image. The precision of the Baseline method is calculated based on the Euclidean distance between the 278-dimensional feature vector of the query image and the images in the database.

Fig. 3. Some Samples in the Corel Photo Gallery.

Fig. 4. Some Samples in the Image Dataset SIMPLIcity.

[1] https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval (Download lúc 6:32 AM ngày 25/12/2016)

B. Execute Query and Evaluation

In the experiment, the proposed method is compared with five image lookup methods using different distance metrics: (1) Euclidean; (2) improved Euclidean: weighted Euclidean metric of each feature dimension; (3) Xing: improved Euclidean distance function and weight matrix, which is the solution of the convex optimization problem; (4) RCA: the RCA distance metric improved from the Mahalanobis distance [8]; and (5) MCML: MCML distance metric is improved from Mahalanobis distance whose weight set is the result of data transformation with label constraints. In the experiment, our proposed method (ODLDA) performs retrieval on the deep feature set combined with the optimal Mahalanobis distance function. Results were obtained over three scopes of 50, 100, and 150. Note that the value of each scope is the top of the images returned by each retrieval loop. The reason we take these three scopes is that users often don't have the patience to choose more than 150 responses.

The average precision of the methods is shown in Table I. In this table, we find that the method using the original Euclidean metric has the lowest precision. The three methods, including Xing, RCA, and MCML, have similar precision. Our proposed method has the highest precision.

The average precision-scope curves of the Improved Euclidean, Xing's distance, RCA, MCML and ODLDA are shown in Fig. 5. These are the precision values of the top 50, 100, and 150 images after the first two iterations of feedback. In addition, in Fig. 5, we also draw the Baseline's precision for comparison purposes. According to these results, our proposed method outperforms better than the remaining methods. Thus, on two benchmark data sets, the precision of our proposed method is higher than that of the Improved Euclidean, Xing's distance, RCA, MCML and ODLDA methods. This reinforces that the idea of the proposed method is very effective.

TABLE I. COMPARISON OF AVERAGE PRECISION OF METHODS IN THE 50, 100, AND 150 SCOPES ON THE COREL DATASET

Method	Average precision by scopes		
	50	**100**	**150**
Euclidean	0.2887	0.3065	0.3199
Improved Euclidean	0.3135	0.42658	0.4846
Xing	0.3324	0.47658	0.5125
RCA	0.3424	0.48058	0.5015
MCML	0.3328	0.47958	0.4925
ODLDA	**0.4836**	**0.5065**	**0.5199**

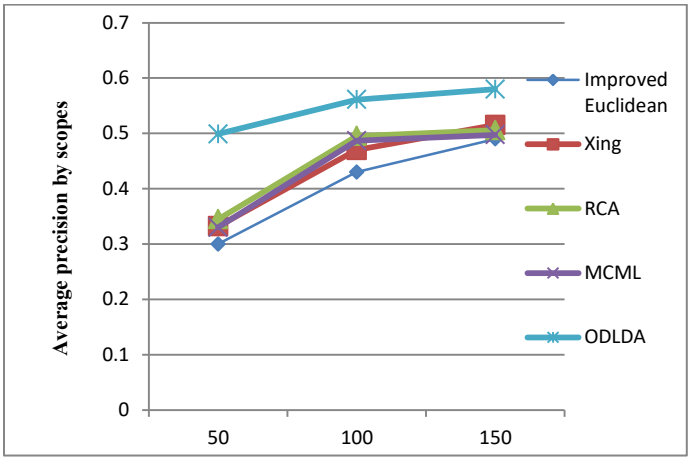


Fig. 5. Comparison of Average Precision of Methods in the 50, 100, and 150 Scopes on the SIMPLIcity Dataset.

V. CONCLUSION

This paper presents the ODLDA method, an effective image retrieval technique for improving the performance of multipoint image retrieval systems. ODLDA effectively exploits the user's information through the relevant and irrelevant sample set, which performs learning an optimal projection to separate irrelevant images and narrow the distance of related images. The proposed method finds the optimal weight matrix of the Mahalanobis distance function and uses this improved distance function to rank the entire database image set and return the result set to the user. Experimental results on two databases have proven that ODLDA provides much greater precision than the Euclidean, improved Euclidean, RCA, and OASIS methods.

ACKNOWLEDGMENT

The author gratefully acknowledges the many helpful suggestions of the anonymous reviewers during the preparation of the paper. This research was supported by the research support program "Offer for senior researchers in 2021" under grant no. NVCC02.01/21-21 and "Improve the efficiency of content-based image retrieval through metric learning" under grant no. VAST01.07/19-20.

REFERENCES

[1] Andre B, Vercauteren T, Buchner AM, Wallace MB, Ayache N (2012). Learning semantic and visual similarity for endomicroscopy video retrieval. IEEE Transactions on Medical Imaging. 31(6):1276–88.

[2] Ruigang Fu, Biao Li, Yinghui Gao, Ping Wang, (2016). Content-Based Image Retrieval Based on CNN and SVM, 2nd IEEE International Conference on Computer and Communications, 638-642.

[3] Monique Laurent, Franz Rendl, "Semidefinite Programming and Integer Programming", Report PNA-R0210, CWI, Amsterdam, April 2002.

[4] A. Globerson and S. Roweis. Metric learning by collapsing classes. Advances in Neural Information Processing Systems, 18:451, 2006.

[5] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. Advances in Neural Information Processing Systems, 18:1473, 2006.

[6] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In ICML, pages 11–18, 2003.

[7] J. Wan,D. Wang,S. C. H. Hoi, and et al, "Deep learning for content-based image retrieval: A comprehensive study," ACM International Conference on Multimedia,pp. 157-166,2014.

[8] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, Learning a Mahalanobis Metric from Equivalence Constraints, in Journal of Machine Learning Research (JMLR), 2005.

[9] C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. IEEE Trans. Pattern Anal. Mach. Intell., 24(9):1281–1285, 2002.

[10] Q. Liu, H. Lu, and S. Ma. Improving kernel fisher discriminant analysis for face recognition. IEEE Trans. on Circuits and Systems for Video Technology, 14(1):42–49, 2004.

[11] G. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. John Wiley, 1992.

[12] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In Proc. IEEE NN for Signal Processing Workshop, pages 41–48, 1999.

[13] M. Guillaumin, J. J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In ICCV, pages 498–505, 2009.

[14] J.-E. Lee, R. Jin, and A. K. Jain. Rank-based distance metric learning: An application to image retrieval. In CVPR, 2008.

[15] A. S. Mian, Y. Hu, R. Hartley, and R. A. Owens. Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. IEEE Transactions on Image Processing, 22(12):5252–5262, 2013.

[16] Z. Wang, Y. Hu, and L.-T. Chia. Learning image-to-class distance metric for image classification. ACM TIST, 4(2):34, 2013.

[17] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS, 2005.

[18] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. Journal of Machine Learning Research, 11:1109–1135, 2010.

[19] D. T T Quynh, N H Quynh, PV Canh, NQ Tao, An efficient semantic – Related image retrieval method, Expert Systems with Applications, Volume 72, pp. 30-41, 2017.

[20] E. Xing, A. Ng, and M. Jordan. Distance metric learning with application to clustering with side-information. In NIPS, 2002.

[21] Flickner, M., Sawhney, H., Niblack, W., et al., (1995). Query by image and video content: The QBIC system. IEEE Computer Magazine 28 (9), 23–32.

[22] A. Pentland, R. W. Picard, and S. Sclaroff (1996). Photobook: content-based manipulation for image databases.International Journal of Computer Vision, 18(3):233–254.

[23] M. Ortega-Binderberger and S. Mehrotra (2004). Relevance feedback techniques in the MARS image retrieval systems. Multimedia Systems, 9(6):535–547.

[24] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N.Yianilos (2000). The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. IEEE Transactions on Image Processing, 9(1):20–37.

[25] C. Carson, S. Belongie, H. Greenspan, and J. Malik (2002). Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(8):1026–1038, 2002.

[26] J. Z. Wang, J. Li, and G. Wiederhold, (2001). “SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 23, no. 9, pp. 947-963.

[27] A. S. Razavian,H. Azizpour,1. Sullivan,and et aI,"Cnn features offthe-shelf: An astounding baseline for recognition," IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 512-519,2014.

[28] J. Donahue, Y. Jia, O. Vinyals, and et aI, "Decaf: A deep convolutional activation feature for generic visual recognition," Computer Science. vol. 50,pp. 815-830,2013.

[29] A. Babenko, A. Slesarev, A. Chigorin, and et aI, "Neural codes for image retrieval," vol. 8689,pp. 584-599,2014.

HADOOP: A Comparative Study between Single-Node and Multi-Node Cluster

Elisabeta ZAGAN[1], Mirela DANUBIANU[2]
Stefan cel Mare Unive1rsity of Suceava, 720229, Romania
Integrated Center for Research, Development and Innovation in Advanced Materials
Nanotechnologies, and Distributed Systems for Fabrication and Control (MANSiD)
Stefan cel Mare University, Suceava, Romania

***Abstract*—Data analysis has become a challenge in recent years as the volume of data generated has become difficult to manage, therefore more hardware and software resources are needed to store and process this huge amount of data. Apache Hadoop is a free framework, widely used thanks to the Hadoop Distributed Files System (HDFS) and its ability to relate to other data processing and analysis components such as MapReduce for processing data, Spark - in-memory Data Processing, Apache Drill - SQL on Hadoop, and many other. In this paper, we analyze the Hadoop framework implementation making a comparative study between Single-node and Multi-node cluster on Hadoop. We will explain in detail the two layers at the base of the Hadoop architecture: HDFS Layer with its deamons NameNode, Secondary NameNode, DataNodes and MapReuce Layer with JobTrackers, TaskTrackers daemons. This work is part of a complex one aiming to perform data processing in Data Lake structures.**

Keywords*—Hadoop; HDFS; single-node cluster; multi-node cluster; namenode; secondary namenode; datanodes; jobtracker; tasktrackers

I. INTRODUCTION

Before the term Big Data, appeared about 15 years ago, there were few possibilities to process terabytes of data sets or higher. Once data generation capacity has increased, needs to store and process this data appeared. The large volume of data has brought with it the need for significant changes in the architecture of storage and processing systems. Over the years, several hardware high-end solutions have been found and implemented, which were smart but very expensive.

Starting from 2003, Google developed a new technology having two main components: Google File System (GFS) [1] and MapReduce [2].

Google created a platform on which multiple data management applications could be implemented. At the same time, Doug Cutting had begun working on a new open-source implementation based on the ideas suggested by Google, so Hadoop was born. [3].

Hadoop is a distributed processing software framework that can process both small and large volumes of data across clusters of computers. However, it is recommended for large data sets, because it is able to scale-up from a single server to hundreds.

Over time, Hadoop has consolidated its position through some advantages, such as:

- All data are accessible, therefore is no need to archive/clean data before storage because Hadoop allows increasing storage space without limits.
- Is a scalable architecture that allows the addition of new clusters/servers to the original architecture without limits. There is no need to upgrade the server but you can simply add new clusters to increase storage capacity.
- Is a robust and easy-to-use architecture that can be easily configured by modifying the config file which is usually an excel file.

Hadoop is based on the Hadoop Distributed File System (HDFS) that allows data to be distributed to multiple nodes. Thus, data are read in parallel and the time required for this operation is substantially reduced. Another important specific feature of Hadoop is that it is based on the "write once and read many times" technology. Even if the writing process will take longer, the reading process makes an important contribution to reduce the read/analyze data time.

In the following, we will analyze Hadoop architecture and we will compare its two ways of implementing: Single-node cluster and Multi-node cluster. This work is structured as follows. After the brief introduction discussed in Section I, Section II addresses some related works, and Section III presents the Hadoop architecture. Section IV outlines the two ways of setting up Hadoop and are highlighted the differences between them. Section V is intended for final conclusions.

II. RELATED WORKS

According to the latest research in the field on Hadoop, it can be said that this is a robust and easy-to-use architecture, it can be easily configured by modifying the *config* file. It is also a scalable architecture that allows new clusters/servers to be added to the original architecture without the need for further upgrades.

In [4] the authors present the Hadoop Single-node cluster installation and setup, and also the required software used in their exemplification. They implemented their cluster on Hadoop Version 2.7.2 using *Jdk- 1.7* as java version.

Based on the definitions found in specialty articles, HDFS and MapReduce is a scalable and error-tolerant model that hides all complexities for Big Data analysis. Thus, in Article [5] Hadoop and its components, which include MapReduce and HDFS, are discussed in detail.

The paper [6] presents a methodology based on the reference analysis to guide the implementation of the Hadoop cluster. The authors analyzed local and cloud architectures using centralized and geographically distributed servers. The results of the research done by the authors show that the methodology can be applied dynamically based on different architectures. At the same time, the authors state that the acquired knowledge can be used to improve the data analysis process by understanding Hadoop architecture.

In [7] the authors give a brief description of the Hadoop ecosystem and several components like Pig, Hive, Mahout, Qooop, Hbase, Flume. The authors also implemented a Hadoop-based platform for the analysis of collected tweets. The obtained results are then transferred to graphic charts.

III. Hadoop Architecture

Hadoop is based on master/slaves architecture. Thus, in such architecture, there is a master and more slaves, where the master manages all Hadoop activities by hosting the NameNode and the JobTracker/MapReduce, and the slaves are intended to store the data hosting the DataNodes/HDFS and the TaskTracker/MapReduce.

The following components are parts of Hadoop:

1) HDFS layer: Storage layer based on a master/slave architecture.

a) NameNode (master daemon).

b) Secondary NameNode.

c) DataNodes (slave daemon).

2) MapReduce layer: Data processing layer-based also on a master/slave architecture.

a) JobTrackers (master daemon).

b) TaskTrackers (slave daemon).

HDFS and MapReduce is a scalable and fault-tolerant model that manages all complex processes of BigData analytics. Hadoop distributed file system - HDFS is a Java-based distributed file system that allows the storage of a large volume of data across multiple nodes in a Hadoop cluster. Using HDFS provides multiple benefits such as distributed storage, distributed and parallel computation and horizontal scalability. On the other hand, MapReduce is a programming framework that allows performing distributed and parallel processing on a large volume of data in a distributed environment. MapReduce framework uses JobTracker and TaskTracker to monitor and execute tasks.

Fig. 1 shows the master/slave Hadoop architecture graphically representing the two essential components and their interconnections.

NameNode stores system metadata and manages clients' requests. When data is stored in HDFS, the NameNode is the one that will keep all the storage information. NameNode is actually the master in the master/slaves architecture, which is why it is also called master-node. NameNode is the main node of the architecture, and if this node fails then the entire Hadoop system will fail. Physically the master node will need the best hardware resources because that's where all the metadata will be kept.

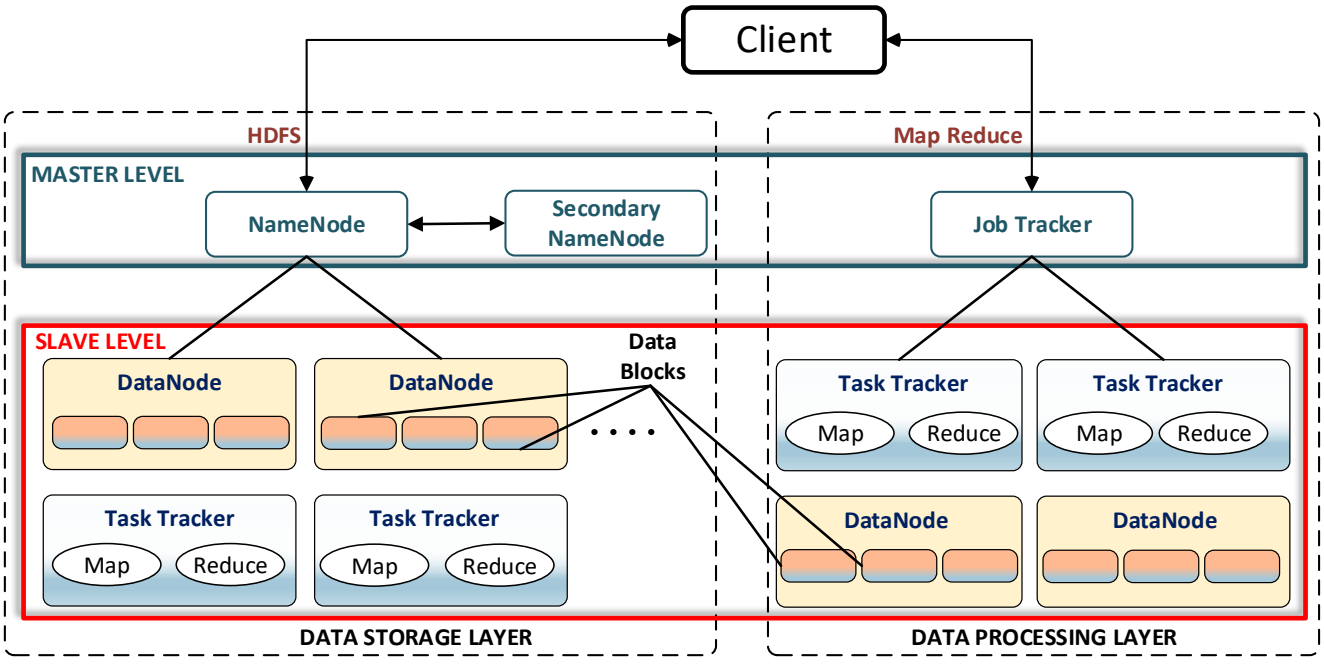


Fig. 1. Hadoop Master/Slave Architecture.

Metadata are data sets that have crucial information about system files: lists of files and folders in HDFS, lists of blocks and where they are stored, information about data permissions (read, write, run), access times, etc.

There are two different types of metadata: FSimages and EditLogs. EditLogs hosts all logs generated by the master node throughout its operation. Any actions and changes that occur in HDFS will be saved in this EditLogs. FSimage are metadata that will be saved in the RAM of the master node and hosts the image of the system files.

Let's assume that the master node has failed, so we need to restart the entire system. When restarting, the master node must reload the image from FSimage and upload all data saved in EditLogs. The larger the log files is, the longer the restart time will be, it can last even tens of minutes, which is not the desired time to wait for a restart. To faster this restart procedure, a second master node named Secondary NameNode has been created, which does nothing but query the logs in EditLogs and update FSimage with the new data. Thus, if the master node fails, it will access the updated FSimage files, reducing the restart time. It is very important not to confuse the function of the Secondary NameNode as a true copy of the NameNode. The Secondary NameNode serves only to update the FSimage in order to reduce the restart time of the main master node in case of a failure. In other words, Secondary NameNode is used to create restart points of the main master node.

In Hadoop, the data are stored in blocks, and these blocks are saved in DataNodes also called SlaveNodes. These DataNodes send information signals once every 3 seconds to the master node, so NameNode can be seen as the heart of the entire system that pulsates and receives information continuously. These signals constantly sent, inform the master node that the data is ready at any time to be accessed. All system operations that take place on a DataNode will be constantly sent to the master node.

As we already saw, within Hadoop the data is stored in HDFS files, and the data is broken into individual blocks. The standard size of a block in Hadoop is 128Mb, but this size can be easily changed from the config file. Therefore, every file that is sent to DataNodes, will be saved in individual blocks of 128Mb. If, for example, there is a 130Mb file that needs to be saved, then it will occupy 128Mb in one block and in another block only 2Mb. This is how data are stored in blocks at the level of HDFS files in Hadoop. The number of blocks is given by the file size divided by the size of the data block. One of the main features of HDFS is that, once data blocks are stored in DataNodes, they will be automatically replicated in different DataNodes to provide fault tolerance. The replication number, in general, should be set up to 3 in the config file.

JobTracker runs on master node level and keeps track of all tasks that are serving DataNodes, known as TaskTrackers. Within this one master and multiple slaves architecture, we will have a single JobTracker and several TaskTrackers. JobTracker is an essential daemon for MapReduce and his job consists in receiving client requests for MapReduce, communicate with the NameNode to determine the location of the data, finds the best TaskTracker nodes to execute tasks based on the data locality and node availability, monitors the TaskTrackers independently, inform the client about the job status. When JobTracker is down, HDFS will continue to work, but MapReduce cannot be started because of the MapReduce jobs that are halted.

TaskTrackers are designed to keep track of all actions that are running on DataNodes, sending the information back to JobTracker. TaskTrackers will be assigned by JobTracker to execute Mapper and Reducer tasks on all DataNodes (Fig. 2). TaskTracker will constantly inform the JobTracker about the progress of the tasks in execution. When a TaskTracker no longer responds, then JobTracker has the ability to assign the task performed by the faulty TaskTracker to another node.

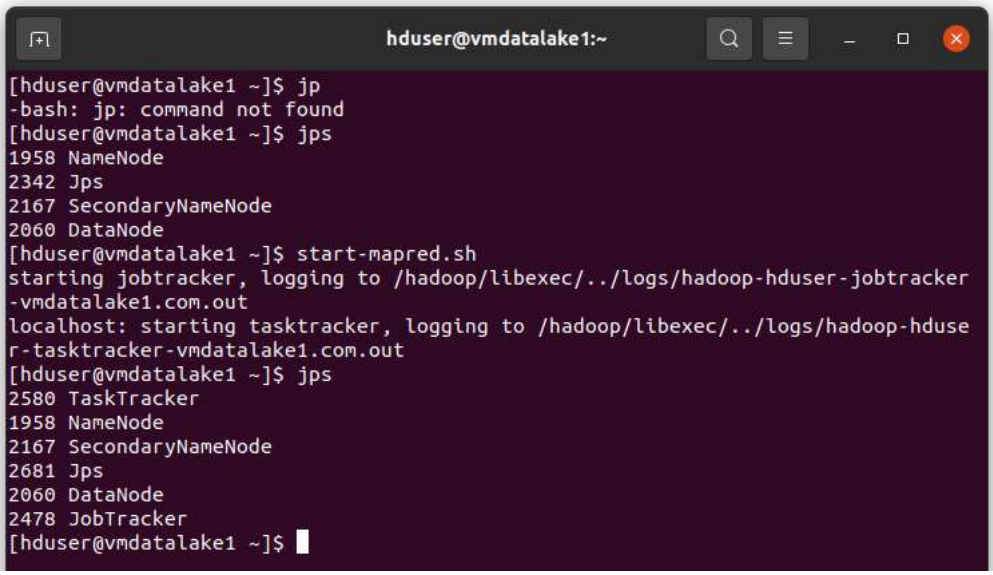


Fig. 2. Single-Node Cluster Daemons.

IV. HADOOP: SINGLE-NODE VERSUS MULTI-NODE CLUSTER

There are two ways to setup a Hadoop architecture: Single-node cluster and Multi-node cluster. In the following, we will perform a comparative study between these two approaches.

For the practical implementation, we used a PC with Ubuntu Desktop 20.04.1 LTS operating system on which we installed Hadoop 3.30. We used VMware Workstation 16 Pro to create virtual machines on which we installed an older version of Centos - Centos-6.3 which is known as a very stable version for this purpose. The purpose of this work is not to explain how to setup and configure, which can be found in [8], [9], but is intended to exemplify the two methods of implementing Hadoop, and to highlight the differences between them.

A. Single-Node Cluster

Single-node cluster is a method of implementing and setting all daemons on a single virtual machine. This setup method is generally used for the study and the test phase, or in environments with fewer data, but in this case, maybe Hadoop is not the most recommended technology to store data. If the volume of data is not large enough then you can hardly notice the main advantages of Hadoop from the first place. After installing, configuring, and starting ssh processes of a Single-node cluster, launching the *jps* command, which is a java virtual machine process status tool, we can see the status of all Hadoop daemons like NameNode, Secondary NameNode, JobTracker, TaskTracker, DataNodes that are currently running on the machine. As we can see in Fig. 2, they all run on the same virtual machine that we name vmdatalake1.

B. Multi-Node Cluster

Setting Hadoop in Multi-node cluster involves the use of more than one virtual machine VM. Each data node runs basically on a different VM. This type of configuration is used in organizations to analyze BigData. In the real environment when you are dealing with petabytes of data, this data must be distributed in hundreds of machines to be processed in real-time. Next, we can see the implementation of a Multi-node cluster using the same system environment as in Single-node cluster. We configured six virtual machines (Fig. 2) by cloning the VM from the Single-Node cluster, each of which we set it up after for specific purposes.

By running the same java jps command on each VM you can see which daemons are running on each dedicated VM, where NN corresponds to the VM dedicated to the NameNode, JT to JobTracker, SNN to Secondary NameNode, DNn to DataNodes: DN1, DN2, DN3. Fig. 3 is a capture launched from another computer connected to the same LAN over the MobaXterm app by using Write commands on all terminals.

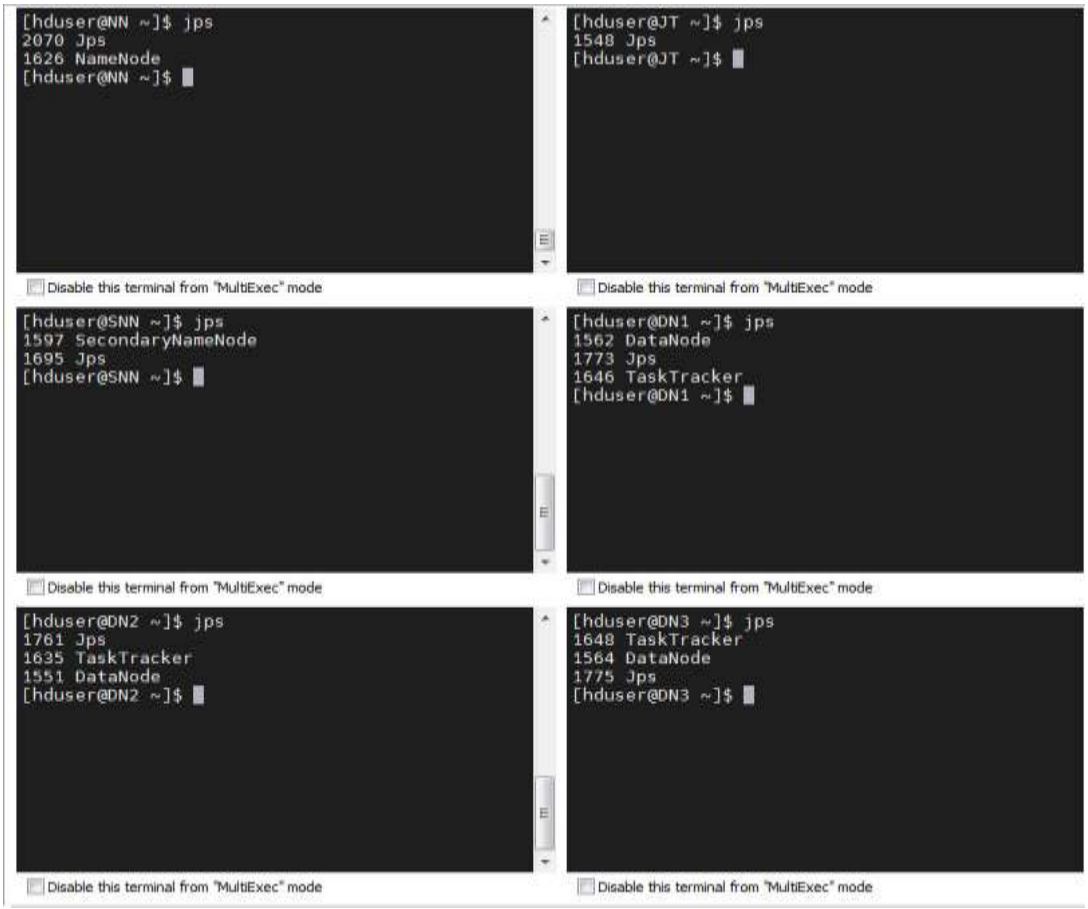


Fig. 3. Multi-Node Cluster Daemons.

By running the command: *hadoop dfsadmin -report* [10] under the NameNode as the HDFS superuser *hduser* in our case, it can be checked if everything is working well. This is one of the first commands that it should be learned as a Hadoop admin, it is one of the commands used frequently to obtain a full report about your data nodes.

Fig. 4 is a report capture over the VMs configured above and we can see: what is the cluster capacity, how much DFS is used and how much is remaining in every single machine, what is the entire cluster capacity, what are the corrupted blocks.

Table I shows a comparison between Hadoop Single-node cluster and Multi-node cluster.

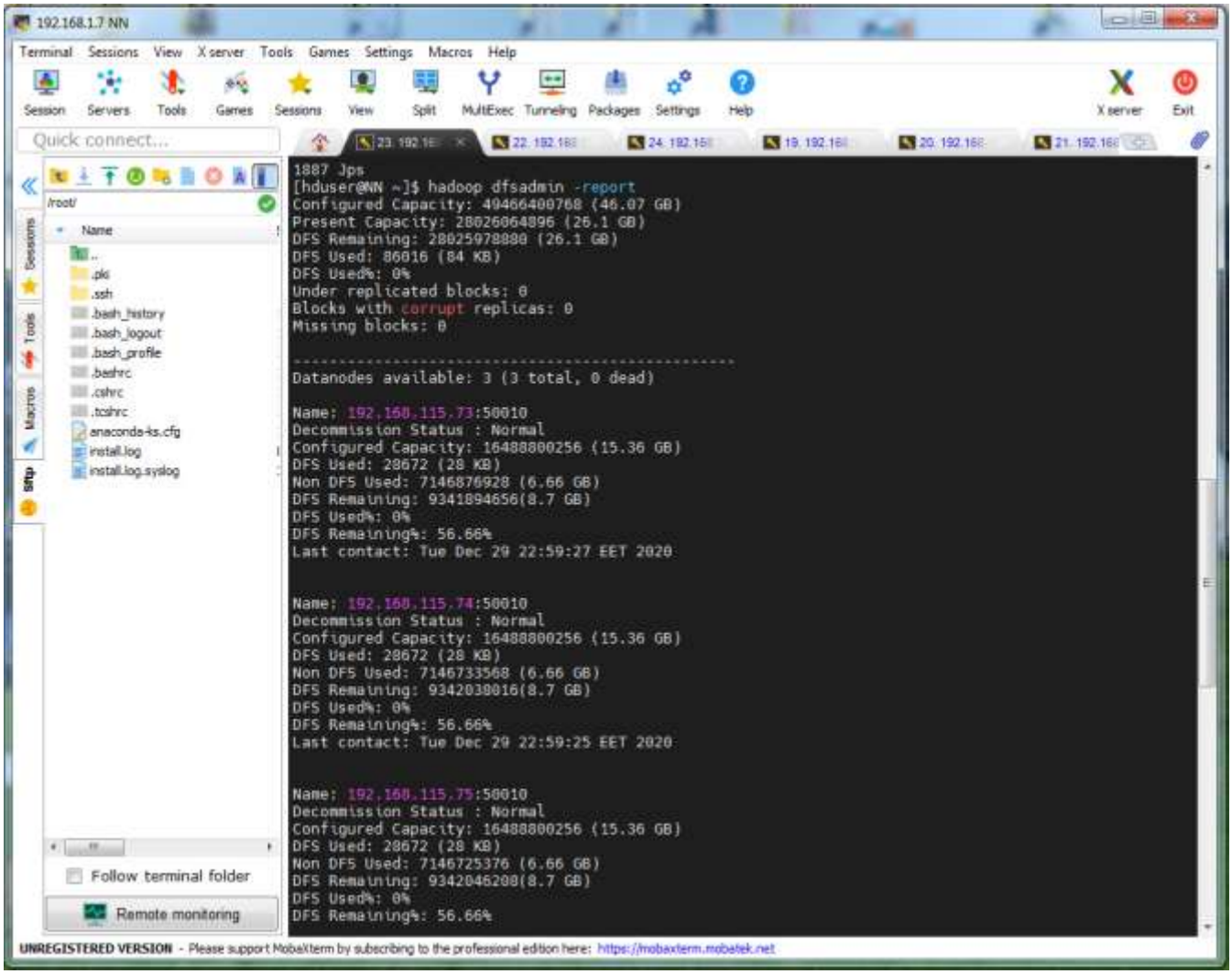


Fig. 4. Hadoop Multi-node Cluster Report.

TABLE I. SINGLE-NODE CLUSTER VERSUS MULTI-NODE CLUSTER

Single-node cluster	**Multi-node cluster**
Hadoop is installed on a single machine or data node.	Hadoop is installed on multiple data nodes runging from a few to hundreds of nodes on a LAN network.
All Hadoop daemons NameNode, DataNode, Secondary NameNode, JobTracker, TaskTracker runs on one single machine.	In distributed mode, NameNode, DataNode, Secondary NameNode, JobTracker, TaskTracker run on different machines.
The replication factor is one in the Single-node cluster.	In Multi-node cluster the replication factor will be greater than one and it should be installed in more than one machine.
Predominantly used in the testing and study phases. It is used also in basic tests.	Used within BigData-type high-volume organizations.
Used to run simple MapReduce processes and HDFS operations.	Used for complex computational requirements such as BigData analytics.

V. CONCLUSIONS

Organizations around the world have found in Hadoop a simple and highly efficient model that works well in the distributed environment, Hadoop becoming more and more used. Applications running on Hadoop clusters are always improved and constantly evolving to meet all market requirements.

In this article we presented a brief description of the Apache Hadoop framework along with its main components, highlighting the major advantages that this BigData storage and analysis system brings. The research is concluded by presenting the two ways of the practical configuration of the Single-node cluster and Multi-node cluster in Hadoop and by a comparison of these implementations carried out in practice using Hadoop 3.0. This work is the initial step of a complex project that aims to contribute to data processing in Data Lake structures.

ACKNOWLEDGMENT

This work is supported by the project ANTREPRENORDOC, in the framework of Human Resources Development Operational Programme 2014-2020, financed from the European Social Fund under the contract number 36355/23.05.2019 HRD OP /380/6/13 – SMIS Code: 123847.

REFERENCES

[1] http://research.google.com/archive/gfs.html (Accessed Nov. 2020).

[2] http://research.google.com/archive/mapreduce.html (Accessed Nov. 2020).

[3] Garry Turkington, Gabriele Modena, “Big data con Hadoop”, May 2015, ISBN: 9788850333431.

[4] Shah, Ankit & Padole, Dr. Mamta. (2019). Apache Hadoop: A Guide For Cluster Configuration & Testing. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 7. 792-796. 10.26438/ijcse/v7i4.792796.

[5] Rehan Ghazi, Durgaprasad Gangodkar, “Hadoop, MapReduce and HDFS: A Developers Perspective”, Procedia Computer Science, Volume 48, 2015, Pages 45-50, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.108.

[6] Correia, R.C.M.; Spadon, G.; De Andrade Gomes, P.H.; Eler, D.M.; Garcia, R.E.; Olivete Junior, C. Hadoop Cluster Deployment: A Methodological Approach. Information 2018, 9, 131.

[7] Can Uzunkayaa, Tolga Ensaria, Yusuf Kavurucub, “Hadoop Ecosystem and Its Analysis on Tweets”, Procedia - Social and Behavioral Sciences, Volume 195, 3 July 2015, Pages 1890-1897.

[8] Oshin Prem, “Installation and Configuration Documentation”, Sep 2017, (Accessed Dec. 2020) https://readthedocs.org/projects/doctuts/downloads/pdf/latest/.

[9] https://www.edureka.co/blog/hadoop-tutorial/#HadoopFeatures (Accessed Dec. 2020).

[10] Eric Sammer, “Hadoop Operations”, September 2012, Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Technology in Education: Attitudes Towards using Technology in Nutrition Education

Asrar Sindi[1]
Faculty of Humanities and Social Science
Newcastle University
Newcastle, United Kingdom

James Stanfield[2]
School of Education
Newcastle University
Newcastle, United Kingdom

Abdullah Sheikh[3]
Department of Computer Science
Taif University
Taif, Saudi Arabia

***Abstract*—Digital technologies have influenced how teachers conduct the daily practice, and students learn in classrooms. In addition, technology is increasingly being deployed in the classroom environment via a combination of kinesthetics', visual's and auditory approaches. This paper aims to investigate teachers' and students' attitude towards using technology in nutritional education. Then, it discusses the impact of online games to enhance nutritional education of students. After that it will discuss the implication and findings of applying learning games to the curriculum from both teachers and students perspectives.**

***Keywords*—Technology; application; online games; nutrition; education**

I. Introduction

During the recent years, the education landscape has transformed drastically. Accordingly, digital technologies have entirely revolutionised the manner in which teachers instruct, and students learn in the classrooms [1]. The copying of sentences on the blackboards and the habitual use of textbooks are no longer the only approaches utilised in ensuring learners are engaged. Currently, the world is highly dependent on technology every day. Technology is increasingly being deployed in the classroom environment via a combination of kinaesthetic, visual, and auditory approaches [2].

knowledge in an exciting and interactive manner [3]. For instance, the usage of whiteboards or classroom clickers enables educators to vary their teaching techniques while reinforcing the learning materials.

Technological devices have enhanced cooperation and communication within and beyond the classroom environment. McKnight et al. [4] note that learners can share an opinion on various learning social networks, develop multimedia presentations, undertake research and surveys, attend webinar sessions, or even take part in live discussions with other learners positioned in different geographical areas [5].

Nutrition-associated health conditions like hypertension, obesity, and diabetes that are prevalent from children to adults, are increasing the importance of nutrition education [6]. The outcome of such conditions involves physical discomfort, low self-esteem, negative impact on the overall social health, poor physical health, increased healthcare costs, poor academic outcomes, and a higher likelihood of poor health during adulthood. Thus, there have been calls to employ various prevention strategies, including deploying nutrition education among children and adolescents [7]. Habits like eating are shaped at a young stage in life and schools provide an environment that assist in shaping what people eat and creating awareness of its importance, leading to proper growth and development of students [8].

Researchers such as Adams et al. [9] assert that schools are believed to be the most effective environment for preventative nutrition studies. Also, they argue that nutrition education curricula offered in learning institutions have positive implications, including behavioural and psychosocial effects as well as physiological impact. In addition, various stakeholders guided by evidence-based research, are advocating for comprehensive nutrition and public health education throughout the learning process [10]. In nutrition education, technologies are being utilised, McEvoy et al. [11] revealed that teachers use hands-on activities from moderate to a high degree in educating students about nutritional concepts. In another study, students and educators reported that computer applications could be beneficial in nutrition education [12].

An investigation made by Yang et al. [13] concerning the importance of technology in nutrition education revealed that the use of technologies in learning at schools are increased such as using such as video, the internet, and computer-assisted programs. Teachers argued that technology improves discussion, collaborative learning, as well as simulations, practices, and role-playing, thus enhancing the learning process and experience [14].

A positive attitude towards technology is broadly recognised as an essential prerequisite for its effective deployment in learning. On the other hand, Rosi et al. [12] demonstrated that a negative attitude could be a detriment to the utilisation of Information Communication Technology (ICT) in schools. Attitudes toward ICT among learners and teachers are influenced by a number of factors such as level of technological capability, support, encouragement, confidence, availability, training, and policy [11]. Strasburger [14] notes that nutrition educators and learners are prepared to integrate technology into their classrooms effectively. Rosi et al. [12] claim that the educators' capability and choice to incorporate technology into their teaching is motivated by several factors. For instance, some educationalists may lack adequate technical abilities for interfacing with technological platforms. Others have been reported to lack technological self-efficiency or integration self-efficiency [14]. In some cases, technical as well as administrative support has been reported to be missing, making it challenging for teachers to obtain finance in order to attend training courses [13].

Accordingly, this paper aims to examine the students' and teachers' attitudes towards utilising technology in nutrition education in primary schools.

This paper organised as follows. Section II gives an over view of the features that involved to develop learning enhance education. Section III serves the research literature review to present the impact of using online games to develop students awareness of nutritional concepts, then Section IV explains the research methodology. Then, Section V discusses the research results and analyses the teachers and students attitudes towards the technology. Finally, Section VI is discussing the research finding, and Section VII concludes explaining the need of developing online games in the purpose of enhancing nutritional learning.

II. Background

This section serves as a background and a general view of the features that involved to develop learning enhance education.

Deaths throughout the world could result in by Chronic nutrition related conditions, hunger, and malnutrition [15]. Recently, there is a reduction in engagement in physical activities among adolescents and children. Moreover, their eating habits have transformed from traditional food to snacks that are highly unhealthy. According to Vidgen [16], nutrition has a close correlation with mental and physical health. Previous studies concerning the role of school education in shaping healthy eating choices and behaviours among learners have shown positive outcomes. Thus, a nutrition education which models healthy eating lifestyles for students and adopts nutrition programs based on the specific needs of the learners is vital [17].

A. Significance of the Study

Deploying technology in learning environments, including digital tools like hand-held devices and computers, instils an exceptional experience in both learning and teaching. Technology expands learning materials and course contributions, enhancing 24/7 learning by increasing engagement and motivation while accelerating learning [4]. Furthermore, technology ushers in a novel model of interaction between learners and teachers, boosting the delivery of professional content and resources that assist in personalising learning or customising it. To utilise such benefits for nutrition education and develop a multidimensional technological curriculum, it is vital to comprehend the learners' and teachers' attitudes towards using technology [3]. The outcome of the study will provide insight into the futuristic development of ICT in nutrition education and professional development interventions by scholars and policymakers of the discipline. As new technologies emerge in the educational field, comprehending the perceptions of technology is essential. Furthermore, this research will be valuable for scholars intending to undertake further investigations in relevant fields. It offers insights that will guide researchers adding to the pool of studies concerning the integration of technology in education.

III. Literature Review

The aim of this study is to answer main question of the research: "What is the impact of using online learning games to develop student's awareness of nutritional concepts?" In this paper, past studies concerning the use and impact of technology in teaching and learning, as well as the use of technology in nutrition education will be examined. Thereafter, attitudes towards the deployment of technology in education will be assessed. Furthermore, the section will explore the challenges of using ICT in learning. At the end, games in education will be explored, followed by reviewing games for nutrition.

A. The Use and Impact of Technology

1) Technology in Learning: To provide a context for exploring the attitudes towards the use of technology in education, it is vital to define what constitutes technology. The use of technology in learning and teaching processes, the terminology describing such technology is expanding [18]. The concept is loosely employed to define the various applications of computers and their integration into learning. Technology for education include computer-assisted programs, web-based learning, computer-based education, distance learning, and multimedia [18]. Learning and teaching via technology can be defined as using electronic tools in delivering information and facilitating the acquisition of knowledge or skills. The technologies integrate presentation techniques (the manner in which information is delivered to the students such as video, interactive TV, audio, or multimedia) and delivery techniques (the way information is taught to learners). Additionally, it entails how learners search for knowledge and share it with peers [19].

Despite the different kinds of technology utilised in education, it is commonly agreed that they intend to foster flexibility by offering various learning approaches through electronic devices [18]. While studies concentrate on computer-assisted technologies, there are numerous learning and teaching technologies that are not supported by computers [20]. These teaching technologies include visual documentary show, television, DVD/VCR, overhead projectors, sound systems, models, and tape recordings [21]. In this study, technology will be utilised to refer to various kinds of equipment for teaching and learning. The technologies include computer-based, software, web-based learning, and hardware [18]. Mayer [22] observed that the learning process is improved when pictures and words are integrated to offer an interactive experience to the learners which grasps their attention. Accordingly, learners taught using multimedia applications found it much better than those taught with the conventional method. Su [23] analysed the impact of technology on the performance of students in science subjects with the assistance of educational software. The outcome of the study demonstrated that the application of technology contributed positively to learning and affected the attitudes towards learning science subjects. In the subsequent section, a detailed exploration of literature is delivered concerning the impact of technology on education.

2) Technology in Education: Literature concerning technology in education has witnessed a spirited debate among scholars and theorists regarding the effectiveness of employing

technology to assist learners and teachers in education [24]. Generally, the majority of scholars and theorists have a perception that technology enhances learning when deployed in the educational process. There are two groups with different outlooks concerning the deployment of technology in education [25]. Clark [26] argues against the view that technology use can improve the teaching and learning process. Accordingly, technology is just a way that can assist delivering instructions but does not necessarily influence achievement of learners. Technologies do not involve like the teachers role in learning and teaching, but the instructional technique is the active factor that catalyses learning [3].

Other researchers have similar views like [27]. For instance, Jewitt [28] discovered that distance education that deploys technological tools does not vary significantly from the traditional approaches to teaching. Concerning distance learning, the researchers made a case that conventional and technologically-based strategies are both effective and they can substitute each other depending on the situation. The authors concluded that learners should not solely rely on distance learning since it is costly as compared to the conventional course learning. Furthermore, Nomass [29] argues that learning or education is a process that involves a series of stimulus-response linkages. Considering the different views concerning the relationship between learning or teaching and technology, the significant question among researchers is how technology can enhance learning.

Kirkwood and Price [3] argue that there are appropriate and inappropriate applications of technological resources in the classroom. The appropriate application can effectively facilitate learning while inappropriate usage can obstruct it. For instance, a laptop can be used to research and view learning materials or it can be employed by learners in wrong ways like sharing pictures that are not classroom-related resulting in time wastage. Therefore, technology should tap into the cognitive process of learning [25]. Encoding achieved via visualisation can be accomplished using technology. Part of the role of the student is putting information into memory utilising some visual cue, a mathematical instructor can employ a computer to show 3D images of molecules [28]. Such visualisation is richer and more detailed as compared to trying to draw 2D images on the chalkboard; hence using technology simplifies the learning and remembering process.

Nonetheless, technology is a double-edged sword, and when the instructor does not employ it in the right manner due to poor training in pedagogy, it can be less effective [29]. Being aware when to deploy technology in the classroom can significantly assist learners to comprehend materials [24]. Technological advances like simulations or expert systems can offer experimental learning which cannot be achieved using traditional textbook approaches. However, the learning process does not simply change due to the use of technology. Technology enables one to access more information quickly and efficiently. The learners can use less time looking for information and more time on making decisions. On the other hand, the teachers can grasp the attention of learners easily and instruct them using more than one approach that impacts on more than one sense [29].

Currently, several benefits have been identified in utilising digital learning and teaching approaches. Kelly et al. [24] note that technology can enable learners aged between five to eighteen years to access information and learn in an interactive manner. One of the important benefits is raising the accomplishments of students such as greater control over the learning process, rapid acquisition of knowledge, and better performance in tests and exams. Higgins et al. [30] offer a summary of quasi-experiments concerning the role of technology in raising school accomplishments. Accordingly, there is a positive correlation between technology and educational outcomes. Hall [31] found a positive association between ICT utilisation and achievement in a study undertaken in England regarding design technology, maths, foreign languages, and science. Papastergiou [32] demonstrated a connection between high usage of ICT and improved performance in learning and academic achievement. In another study undertaken in Taiwan by Tamim [33] concerning the impacts of digital technologies and resources on elementary students, 92% of the learners exhibited a positive impact due to the use of computer-assisted learning and teaching.

However, 8% showed negative impacts favouring the traditional approach. Studies have also examined how technology impacts literacy levels among learners. For instance, Archer et al. [34] carried out a meta-analysis to examine the outcomes of prior studies that considered the effect of technology on literacy and language learning [35] [36]. The study, overall, revealed a relatively minimal but positive impact of the use of technology on literacy. In addition, classrooms with small numbers of learners tend to show a significant positive impact as compared to classes with many learners. In a meta-analysis conducted by [30], it was demonstrated that digital learning and teaching enhance writing skills such as spelling or reading. Hess [37] explored the effect of utilising e-books and e-readers in the classroom environment amongst learners from the USA aged between nine to ten. The outcome of the study indicated that there was a significant variance in reading achievement for the students who used e-readers with scores increasing for both boys and girls. Thus, technology use seems to have varying impacts on literacy levels. Lysenko and Abrami [38] examined the deployment of two technological tools including online gaming tools in relation in reading comprehension among learners aged from six to eight years old. The results obtained show a slight improvement in reading comprehension.

According to Jewitt [28], technology in science can be used in taking pictures and presenting background information regarding various aspects of learning. In this way, concepts could be more accessible and easier to learn, as well as facilitating project-based learning which is vital to the learners. In another study, Hsu et al. [39] examined the impact of incorporating self-explanation ideologies into technological tools to facilitate the students' conceptual learning regarding light and shadow. The study entailed students from eight to nine years old. Based on the results, while there was no statistically significant variance in test scores of the control and experimental groups, Hsu et al. [39] observed significant variation in the retention scores, whereby retention involves holding a learner in a given grade instead of moving with other students to the next grade when he/she has not acquired basic proficiency in learning. The experimental group using technology performed better than the control group. Moreover, in a study focusing on the use of technology-enhanced teaching in chemistry by Guven and Sulun [40], a significant variation

was found among the control and experimental group.

Higgins et al. [30] observed that technology could assist secondary school students with relatively low literacy levels. Investigations on knowledge and comprehension in social studies before and after utilising online dictionary and thesaurus indicated that there was a significant improvement in knowledge and understanding. Furthermore, Reed et al. [41] revealed that technology could assist learners to catch up with others. The Phonics programme (a technique utilised to teach children reading and writing in English by mixing English sounds to form words) enables digitally assisted students in improving their spelling and reading. An investigation conducted by Tamim et al. [42] using several studies revealed that word processing could positively impact the writing skills among weaker students.

3) Use of Technology in Nutrition Education: The importance of appropriate diet is similar for adolescents, children, and adults. Nutrition education is perceived as an effective tool that can be provided to people using various approaches at the individual, policy, and community levels [43]. Samieipour et al. [44] defined nutrition education as a combination of various educational approaches intended to facilitate or encourage learners to make healthy choices and nutritional behaviours. Consequently, nutrition education can result in improvements in the health and welfare of students. The school-based nutritional knowledge delivery and acquisition is known to be an effective approach in ensuring a positive nutritional attitude and right habits. Hence, policymakers have attempted to include health promotion educational strategies in education to enhance self-efficiency and alter the behaviours of families and learners. Also, they note that nutritional programs can only be effective when guidance is established based on the attitudes, performance, and knowledge. Kupolati et al. [45] examined the impact of teachers' perception on the school nutrition education and how it influences eating behaviours of students in the Bronkhorstspruit District schools located in South Africa. Results obtained demonstrated that the support for nutrition education among schools was limited undermining the ability of schools to influence healthy eating habits among learners. Thus, there is a need to enhance the educators' capability to model a positive eating habit. Also, they revealed that learners were not entirely unaware of healthy eating, but they had limited ability to influence behavioural changes due to the resource-constrained settings. Furthermore, they argue that to encourage healthy eating habits, it is vital that unhealthy choices of food are discouraged, especially from the food vendors; peer influences should also be avoided.

Previous studies have established the importance of attitude and teachers' role in encouraging healthy eating among students. The teachers' responsibilities include adopting nutritional curriculum and modelling healthy choices [46] [47]. However, little is understood concerning the teachers' attitudes to various aspects regarding nutritional eating. Ercan [48] considers nutrition education as a scientific unit due to the rising obesity in the modern world. Accordingly, healthy nutrition is a vital condition for a healthy life because unhealthy eating habits result in some disease.

According to Cooper et al. [49], the right nutrition is significant to cognitive functioning in adolescence and children attending schools. The authors showed the effects that consuming a healthy breakfast had on the level of student's concentration at school. Cooper et al. [49] compared the performance of learners who did not have breakfast with the learners who ate breakfast. O'Dea and Abraham [50] examined the level of knowledge, attitudes, and beliefs that the physical education and home economics tutors had towards eating disorders and obesity problems. The intent of O'Dea and Abraham [50] was to explore the extent to which teachers and students were informed of their wellbeing. As a result, it was identified that a positive attitude of teachers has positive implications in assisting learners to adopt healthy eating habits. Generally, these kinds of studies have different weaknesses since their results were restricted due to the availability of literature that increasingly focused on young children or adults as there were no previous studies conducted by employing adolescent subjects. Moreover, the study by Cooper et al. [49] was limited by a small sample size that did not offer comprehensive results. Sharma and Rani [51] studied the changes in knowledge of IT professionals after nutrition education was delivered digitally for a month. Thus, the study demonstrated that the provision of nutrition knowledge via technology greatly aids in promoting healthy dietary habits.

In addition, there are nutrition curriculums that are aligned with units such as mathematics, social sciences, languages, and arts. One example is the Dairy Council of California K-12 [52]. Each grade in the program possesses a specific curriculum with ten lessons on nutrition. The lessons are made using a behavioural change mechanism that encourages students to eat healthy. Through this program, children can exercise long-term health related skills such as setting a goal, making decisions, and analysing effects. The plans are fun and simple to use, as well as effective for the education of students in general. These are important in the development of the child, cognitively and physically too. Most of the foods and nutrition programs are designed for the students in school. Nutrition educators must be informed and take part in discussions concerning access to information technology. Due its growing importance, in developing countries, there are trends appearing in technology that can support education about nutrition to students [52]. Such technology can be effectively utilised in both nutritional education and other aspects of teaching.

B. Attitudes Towards the use of Technology

Baturay et al. [53] define attitudes as inclinations, bias, fear, convictions, feelings, preconceived notions, prejudices, and ideas concerning a given matter. In psychology, attitude entails a psychological construct, as well as an emotional and mental entity that inheres within an individual and characterises him/her. Moreover, it is organised by experience and exerts dynamic and directive influence on a person's response to objects or situations. In the context of this research, attitude represents the conceptual value of various technologies in the users' mind but not the benefits of the technologies themselves. The usage of technology in education seems to have led to a conflict between individuals who have a negative and positive attitude towards deploying ICT as a learning and instructional tool [54]. Negative attitudes toward educational technologies have been attributed to lack of confidence, insufficient technical support, or lack of pedagogically driven training concerning technology [16]. To comprehend attitudes regarding ICT deployment in education, Condie and Munro

(2007) argue that it is vital to understand concerns that are an integral part of teachers and learners' attitudes. According to Awan [54], these concerns can be categorised as perceptions, feelings, attitudes, and motivations that learners and teachers experience while utilising technology.

Mumtaz [55] argues that before, during, and after the implementation of a novel model, learners and teachers undergo various psychological phases regarding their concerns towards technology. Such concerns can be categorised into three phases including concerns for the self, management and implementation concerns, and concerns about the effect of technology on teachers and learners. That is to say, educational technologies should have a goal of assisting individuals to become more independent by defining the targeted behaviour, functional reinforcement, selecting self-management approach, teaching the use of self-management technologies, and instilling independence. Awan [54] indicates that the present consensus stipulates that a timely determination of the concerns can be important too, in case a learning institution wants to ensure prosperous implementation of technology.

Meerza and Beauchamp [56] examined the attitudes and critical factors affecting the usage of ICT among undergraduates in Kuwait universities. The study found that language, ICT support, and type of institution impacted on the perception towards technology. When the learners and teachers have a positive attitude towards technology, it is likely to be integrated effectively into the learning process. For instance, Rhoda and Gerald [57] demonstrated that a positive attitude is a prerequisite towards using technology in learning. Some of the indicators of positive attitude include improving the presentation, engaging, and making a lesson interesting. On the contrary, some scholars have found that technology can result in negative attitudes including making lessons less interesting, the concepts taught becoming difficult to understand, obstructing learning, and reducing motivation [58] [59] [60].

Balta and Duran [61] examined the attitudes of teachers and learners towards the deployment of interactive whiteboards in elementary as well as secondary school classrooms. The findings from the study revealed that interactive whiteboards are rated highly by learners and students. It was noted that as the learners mature, their attitude towards the use of interactive whiteboards as a technological tool for learning becomes negative. Enayati et al. [62] investigated the attitude of teachers towards the implementation of technology in education in the City of Babol. The teachers increasingly believe that technology has significant benefits and enhances efficacy and effectiveness in education.

Dogruer et al. [63] examined the attitudes of primary school teachers of English language towards utilising educational technology by administering questionnaires. The study showed that there was a positive attitude especially concerning the impact technology has on knowledge acquisition and improving the achievement of students. Furthermore, a study by [64] examining intern teachers' attitudes towards technology indicated that a positive attitude was found with no significant gender differences. Moreover, the study revealed that there are no considerable variations between the field of teaching or subject. Kabadayi [65] undertook an exploration of preschool teachers as well as part-time teachers to determine their attitudes toward multimedia in learning. According to the findings, 75 technologies. Similar findings have been documented in Zanguyi's [66] study examining educators' attitudes toward the deployment of technology in learning. In another recent study, Seraji et al. [67] examined the attitude towards ICT usage amongst teachers originating from various institutions in Mazandaran. The sample involved 62 female and 38 male teachers, and the findings demonstrated that there was a statistically significant correlation between the attitude of teachers and their experience with technology. When the teachers have a positive attitude, technology is perceived positively. Moreover, Seraji et al. [67] discovered that there was a statistical relationship between the teachers' tenure and age with their attitude towards ICT. That is, younger teachers have a considerably more positive attitude towards the use of technology as compared to older teachers.

Chow [68] indicates that age does not determine the teacher's attitude toward technology. However, the extent to which an educator is comfortable and has comprehensive training on technology, directly affects educator's perception towards it. Thus, [68] highlights some of the challenges that may make teachers have a bad attitude towards technology: lack of comfort with ICT, the belief that technology does not assist students to learn, lack of interest, inadequate training, and lack of access to technology. A study by Albirini [69] examined the attitude towards the use of ICT for both teachers and learners. The investigation examined perception, performance, and motivation as well as participation in ICT classrooms. The study's results revealed that there is no significant difference in using modern technology in schools between the learners and educators. Nonetheless, it was noted that online resources help students in learning at their own speed. Moreover, ICT resources motivate teachers and learners as they are interesting and interactive, making teachers and learners value ICT as an effective learning tool [70].

Al-Emran et al. [71] indicate that since the learners are generally on task (use technology to learn and research), they are likely to show positive feelings towards the use of technology or computers as compared to doing their work using conventional approaches. Furthermore, the quantity of non-tasked oriented habits reduces considerably in computer classroom sessions due to the use of multimedia tools for spelling or reading. Hence, the use of digital video as a learning tool can enhance on-task concentration. According to Hwang and Chang [72], the use of modern technology cannot fully replace conventional learning activities but easily complement them. Baturay et al. [53] found that learners with low levels of motivation as well as a feeling of uncertainty about learning can demonstrate positive habits in the course of lessons that utilise computers as compared to the traditional approaches. Shroff et al. [73] revealed several positive perceptions and attitudes among students due to the use of technology, including enhanced class attendance rates, improved cooperation among students in learning, and increased research of learning materials outside the school. Additionally, there is increased self-esteem and selfconfidence among learners with laptops. Tseng et al. [74] found that learners are more motivated as they find ICT more appealing, pleasant, and fun compared to traditional approaches. Thus, the use of technology in the classroom is increasingly viewed with a positive attitude among most teachers and learners due to its interactive and interesting experience.

C. Challenges of Deploying Technology in Learning

According to Al-Fraihat et al. [75], there are several hurdles to overcome when a learning institution wants to deploy technology effectively. Awan [54] claims that studies about the challenges or barriers to deploying technology in schools indicate that attitude is a major issue. The fact that some learners or teachers can resist change due to personal beliefs has been examined as a challenge to implementing technology in learning institutions. A school's organisational structure can cause resistance to successful integration of ICT. Condie and Munro [76] present a framework that offers barriers that limit the use of technology in schools, depicted in Fig. 1.

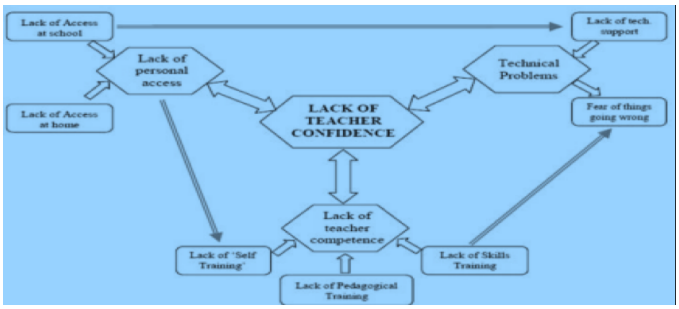


Fig. 1. Barriers to the Deployment of ICT, Source: (Condie and Munro [76].

Hwang and Chang [72] highlight the factors impacting the perceived ease of technology use. The positive factors include regular usage and experience in technology, confidence in utilising technology, owning a computer, and training. The detrimental factors include insufficient access to resources, a challenge to use hardware/software, and lack of technical support [77].

Lack of technological and computer skills is one of the burdens of implementing and having a positive attitude towards the use of technology [75]. Some teachers and students might have basic computer training, technologies keep evolving [75]. Hence, there is a need for regular pedagogical training concerning emerging technologies for them to be updated on recent developments in the field of ICT use in education. Furthermore, training without continuous practice can have minimal impact since people tend to forget. Thus, training is considered a big challenge in deploying technology in learning. At times, before introducing novel technologies, teachers do not receive training regarding their use. Thus, they end up trying to figure out how to use the technology themselves. Osang et al. [78] note that offering training related to the technological features is just one aspect of ensuring successful implementation. The real training related challenge entails training for changes to pedagogical approaches. Blinco et al. [79] assert that technology implementation success rests on the fundamental need that teachers, as well as learners, possess sufficient technical skills to employ technological tools effectively.

Al-Fraihat et al. [75] note that it is normal for all human beings engaged in activities to experience fear of change and resistance since people tend to be more comfortable in holding to the old strategies and processes instead of adapting to new ways. Moreover, the change from old practices to novel ones can be challenging to any entity, be it an elementary institution or higher learning institution [80]. The challenge is attributable to the involvement of massive upfront capital, time, planning, disruptions, downsizing, an increase of workforce, or organisational changes [78].

There is also the fear of being rendered redundant by technology. This is attributed to the fact that people believe that technology minimises the value of teachers once adopted, since they might be dispensed as technology takes up their roles [81]. Hence, teachers might have a negative attitude that can be detrimental towards adopting and using technology as an instructional tool. Moreover, Sethi et al. [82] observed that some institutions lack adequate facilities and resources that have influenced the deployment of ICT in learning institutions. An exploration by [77] indicate that some schools do not have the basic office gadgets such as projectors, computers, binders, printers, and other devices to facilitate the use of technology. In a recent study, Tondeur et al. [80] indicate that it is appalling that some learning institutions do not have access to internet connections. Furthermore, with the spread of technologies throughout the globe, internet security is becoming a major challenge. Specifically, the internet has led to a bad reputation due to fraud, resulting in a fear of electronic transactions. For instance, students can easily outsource their assignments, resulting in loss of trust in the ultimate achievement of learners [77]. Regarding attitudes and beliefs playing a vital role in the adoption of technology, Mayes et al. [83] created a two-phase categorisation that identifies beliefs and attitudes to be the second order challenges that hinder the learners and teachers from utilising educational ICT. The first order challenges involve the obstructions that are external to educators such as infrastructure, training, resource allocation, support, and other software/hardware challenges [82]. The second order barriers are described as the challenges that are internal to the teachers or learners including skillset and confidence that are frequently overlooked [83]. While Osang et al. [78] argue that the first order challenges are important, they suggest that the second order challenges are likely to be more obstinate when infrastructure and resource challenges are alleviated. Table I. shows the first and second order challenges impacting the use of technology.

TABLE I. First and Second-order Challenges [77]

First-order-Barriers (INSTITUTIONAL)	Second-order-Barriers (INSTITUTIONAL)
Extrinsic to the teacher	Intrinsic to teachers (and possibly subconscious or 'private theories')
Lack of access to technology	B eliefs about teaching (and learning)
Insufficient time to plan for integration	Beliefs about computers and technology
Inadequate technical and administrative support	Beliefs about classroom practices and routines
Lack of training	(un) Willingness to embrace change

D. Games in Education

Some of the instructional approaches that employ technology include group discussions, computer-enhanced learning and lectures, or games. Anderson and Barnett [84] studied the use of digital games among learners aged between twelve and thirteen years in America. The study found that digital games

enhanced the comprehension of electromagnetic concepts as compared to students utilising traditional approaches to learn similar concepts. A study by Albirini [69] noted that online resources, as well as games, assist students in learning at their own pace. Also, Condie and Monroe [76] reveal that digital learning and teaching of science made it more interesting, relevant, and authentic to the students.

Online educational games require learners to employ logic, memory, critical thinking, and problem-solving as well as discovery and visualisation [85]. Additionally, the use of such games requires students to manipulate virtual objects through electronic tools and learn complex elements being modelled. Just like other technologies, online learning games have been found to be effective in increasing motivation as well as learners' interest. Nonetheless, the extent to which this translates into effective learning is not obvious [86]. This is attributed to the lack of empirical data because of few systematic investigations on online games and their cognitive effects.

Researchers have attempted to determine the benefits of online games in learning and why the learners find them interesting. According to Young et al. [87], what makes an online game more fun can be explained using psychology and biological functioning of the body. Motivation originates from sensory gratification, adrenaline, engaging environment, roleplaying, taste, the element of fun, and personality. For that reason, games generally motivate players with topics like a survival strategy, building relationships, and roleplaying. Kapp [88] argues that online games can manipulate the unchanged variables based on simulations of natural systems. For instance, in the game SimEarth, learners can observe the impacts of altering the universal oxygen levels or increasing temperature. Thus, the learners can view a perspective from a novel viewpoint [89]. In the Hidden Agenda simulation, students can assume the role of a president in America and learn about sociology, culture, economics, or politics in the process. Young et al. [87] indicate that games are vital in the mental and social development of learners. In an exploration to establish the games loved by learners in the teaching environment, a group of twenty learners played commercial games (Duke Nukem 3D, Simple, Zork Nemsis and Red Alert) [90]. Findings obtained demonstrate that learners prefer 3D adventure (Zork Nemesis) as well as Strategy (Red Alert) games compared with others. Learners ranked game aspects such as memory, problem-solving, logic, and visualisation to be important [90]. These aspects are found in adventure games and are considered essential in the learning process.

With many developers emerging, online game applications are increasing. This has led to several debates concerning the future of online gaming. Some believe it is hard to forecast the future of gaming since significant technological changes are occurring throughout the world, based on how rapidly technology is developing [91]. Not with standing, online gaming will benefit from the continued advancement of online technology. Some researchers believe that the future of online gaming will include an augmented reality as a standard experience whereby players will be in the same room with their adversaries or will be able to see the other players [92].

E. Games for Nutrition

The Institute of Digital Media and Child Development Working Group on Games for Health [93] explored the use of video games in nutrition and health education. According to Taylor [94], games such as My Plate Match could help students to learn about groups of foods and the necessity of each group for healthy consumption. This game mainly targets children aged five to eight years old. It is imperative useful platform for students to gain information about eating habits and a balanced diet. This game takes about 10 minutes and can be used in places where there is internet connectivity. This game application teaches the students to recognize foods that do not fit into any food group, known as extras. Similarly, Granic et al. [95] attribute the effectiveness of video games to the ability of engage with other people who are playing them compared to other media. Studies have found that more than 29% of video game players involve people who are eighteen years or younger. The video games for health are created on platforms that most players are very familiar with such as personal computers, smartphones, game consoles, or web browsers. Moreover, online game is called Mission Nutrition. It has three main tasks; the first task involves critical thinking where the child can determine the kind of foods that have a lot of sugar [96].

The second objective is to look for a snack that provides proteins and carbohydrates while the third task is to test their knowledge about fruits. However, this game ends quickly when learning becomes interesting. The player may search for sugary foods which might promote poor nutrition for the child. Online games are interactive, and it is the role of teachers to proscribe non-interactive media that do not promote health education. Children in schools should learn how to possess cognitive and social skills in the technological world. The main goal of the game is to not only choose the correct answers but also learn about different foods and their effect on the body [94]. Various video games are being created and deployed across a wide array of medical issues, including pain management, human immunodeficiency, cystic fibrosis, and even obesity. Such games are developed for all ages. One of the online games that was developed for use in education was the Immune Attack which was created by the Federation of American Scientists (FAS) [97]. The intent was to teach complex immunology and biology subjects. In the game, a teenager with a distinctive immunodeficiency needs to teach their immune system how to function properly, failure to which he/she risks losing his/her life. The human body acts as the playing field while the immune cells fight viral and bacterial infections with each level featuring a different infection [98]. Another game named Awesome Eats, which is supported by Whole Kids Foundation [99]. The game starts with a chapter that has a "Did You Know?" statement. It allows the students to choose the foods that are not good for their body and those that promote optimum health. After that, ratings are assigned after each level and advances with excellent playing. It is an interactive game for the students who play effectively under a timed challenge. Furthermore, Ship to Shore is also a game that allows students to make choices about the supply of food [100]. It utilises nutrition as a vehicle to also integrate other subjects such as arts, science, and mathematics.

IV. Methodology

This study will answer the main question of the research: "What is the impact of using online games to develop student's awareness of nutrition?" by exploring the following questions:

1) What is the impact of using online learning games to develop students' awareness of nutritional concepts?
2) What are the attitudes of learners and teachers towards the use of online learning games in nutrition education?
3) What are the perceived challenges by learners and teachers towards using online games in nutrition education?

The choice of a research method is an important element because it determines the outcome of the research, how the research questions will be answered, and what needs to be done to ensure that right results have been achieved [101]. A good research method makes it possible to collect sufficient data, analyse it appropriately, and give the right output. Bryman [102] explains that research methods are a way of explaining the beliefs and philosophical understanding of the researcher, and thus are in a position to provide a theoretical background of the research. Therefore, this section will explain the research paradigm and approach taken. Subsequently, methods used, as well as the ontology and epistemology of the theoretical framework of the researcher will be identified. Thereafter, the participants and data collected will be described. The last section will discuss the limitations.

A. Research Paradigm

The research paradigm comes from a Greek word that is used to refer to a pattern [103]. The research paradigm was developed to mean the way people think, and thus became part of the methodological approaches [104]. McCusker and Gunaydin [105] explain that the research paradigm focuses on the views, beliefs, and approaches that one follows. There are three forms of research paradigms, which are the positivist, interpretivist, and pragmatic perspectives.

This research will utilise the positivist research paradigm which is focused on ensuring that the data collected is objective. The positivist research paradigm, according to Brannen [101] separates the knowledge from any other person, because data is believed to be a scientific phenomenon that cannot be influenced by personal opinions. Data collected is presented through facts and figures in this paradigm. The emphasis of the research is often on the objective, which focuses on answering the question "what". Therefore, the quantitative research design is seen to be the most appropriate in utilising the positivist research paradigm. Positivists further believe that the researchers collecting data should separate themselves from the data being collected to avoid bias.

Unlike other methods such as the interpretivist that was not selected for this research, the positivist is highly scientific, objective, and results oriented. The use of the interpretivist paradigm relies on the researcher forming personal connections when collecting data from the respondents. This is because interpretivists believe that the knowledge, data, and information being collected is directly related to and cannot be separated from a person who owns it as [106] explains. Interpretivist information is relayed through the experiences and beliefs of a person [107]. Thus, the method using this paradigm is subjective and highly influenced by personalised connections. Data collected during an interpretivist research approach is usually focused on answering the question of how and why, and could lead to lengthy and in-depth data being collected, thus leading to intersubjective outcomes. For these reasons that relate to the interpretivist research paradigm, the method was not deemed suitable for the research.

The pragmatic research paradigm believes in utilising what works, and in most instances combines the use of positivist and interpretivist research approaches. In this case, it was also not selected for use because the research purely agrees with the use of the positivist paradigm.

B. Research Approach, Design

A research approach can be classified as either being deductive or inductive. A research approach is a way of reasoning, which can help to arrive at a specific or general conclusion about a given subject [108]. This research will utilise the deductive research approach because it involves testing of theories. Landrum and Garza [109] mention that the deductive research approach starts from a specific point of focus such as the selected theory and works towards a more general outcome. This is because the theories being tested at the start involve a hypothesis, and further proving whether they have been proven right or not. The understanding of teachers' and students' attitudes towards technology is an aspect that can be investigated by looking at a tentative theory of whether there is a correlation between two or more variables. This directly relates to the use of the deductive approaches.

This research will adopt a quantitative research design which utilises numerical aspects to collect and analyse data. In other words, a quantitative research design utilises numerical values to represent data. A quantitative approach was adopted as it makes it possible to quantify such aspects as attitude, behaviour, and opinion. The use of surveys has been termed as one of the most effective and objective ways to approach a research question. This research has utilised this method for various reasons that will be justified later in the methodology.

C. Methodology

The factors that affect the selection of a methodology are vast. Thus, the researcher has to understand the current study and consider making a choice that is in line with most of the expected outcomes. The factors affecting methodological selection can be divided into various parts such as practical factors, theoretical factors, ethical factors, as well as the nature of the topic. One of the factors that affect research, as mentioned, is the theoretical factors. Precisely, theoretical factors involve the areas of theory that the researcher can relate to, which involve the validity, reliability, beliefs, and representativeness of the research. The validity and reliability of data are aspects that can be tested to prove its authenticity and realness [110]. In this case, the selection of quantitative methods such as surveys provides one of the easily testable data sets for reliability and validity.

The beliefs of the researcher, in this case, played a big part in choosing the research method. Given that the research

paradigm is positivist and the research approach is deductive, the philosophical perspectives coincide with the choice of a quantitative research method. Representativeness involves whether a research can cover a sample that is appropriate enough to showcase all the characteristics of a given population [106]. The choice of certain methods limits the selection of a wider sample, such as the use of interviews, which can only be done at length to a select number of respondents. However, methods like questionnaires can be administered to a wider population, hence highly representative [109], which is why the method was chosen. Practical factors include time, costs, funding, access to respondents, and personal skills. Large research might be time-consuming because they require lengthy methods of data collection.

This research considered the average time of conducting data collection and analysis before choosing the method. Conducting research is expensive and could be hard to conduct especially if the source of funding is limited. This research was funded by the researcher and was considered to be within acceptable limits. The access to respondents nearly limited the research, but a certain degree of useful data was collected. Concerning ethics, this research was conducted in a manner that complied with ethical guidelines such as confidentiality and voluntary participation. The nature of the topic is also important in choosing what methods to use. For studies related to understanding "what", analysing attitudes, perception, and behaviour based on a direct relationship, quantitative methods would be best placed to allow for a correlation or regression analysis of different variables. When it comes to sensitive topics that explain for example domestic violence, or areas that focus on explaining what or how in research, there is more likelihood of choosing methods like focus groups, interviews, or observations. This research is geared towards the understanding of attitudes and technology, and thus the choice of quantitative methods was the most appropriate.

D. Limitations

Conducting this study in schools requires longer time which could help to obtain better results. Also, using multiple methods such as quantitative together with qualitative. Qualitative such as view of students, teachers, school headteachers, and parents and quantitative such as surveys. Most important is that using online learning games in schools would make teachers notice the difference between using the games and the traditional learning method in students' motivation and comprehension. That would encourage teachers to integrate technology with traditional learning.

V. Results and Analysis

This section presents the analysis of data, which was aimed at facilitating the answering of the overarching question in this study. In the context of the research question which was to investigate the effect or the impact of online games on creating awareness about nutrition among students, it was imperative that the study examines the attitudes of students and teachers towards the use of technology. As such, the researcher prepared both students and teachers questionnaires in an attempt to investigate not only the attitudes of students towards technology in learning about nutritional education, but also the attitudes of teachers, who are an important part of the students' learning process. Nutrition education has become one of the most important and widely studied phenomena, and this has primarily been due to the fact that healthy eating has become an important topic in recent times. The increased awareness in some parts of the world on the need to eat healthy foods and the role of technology has challenged researchers to investigate how technology can promote awareness, and whether or not the awareness can equally cut across people of all ages. This is because young people know very little about healthy eating, partly because of ignorance or simply the fact that they do not care much about it. Nutrition education has largely targeted older people, or even in situations where young people have been targeted, they have hardly received this message with the desired enthusiasm. The analysis in this chapter utilised descriptive statistics, where apart from describing the findings, the researcher interprets the intuitive meaning of these figures, clarifying what it means when respondents reply the way they did.

A. Teachers Attitudes

The attitudes of the teachers towards technology is optimistic, where they believe that the use of technology could go a long way in influencing outcomes even they do not understand how ICT can be used for nutrition curriculum. Moreover, they indicated that they have computers at home and in schools, as a result, they seem quite confident using technology. Teachers, as has been established in this research, embrace technology and highlight its importance in the learning process. The teachers also indicate that online games are an effective form of learning, especially in the context of nutrition education. The teachers also expressed their enthusiasm for using technology such as online learning games supporting having better teaching especially for nutrition education.

B. Students Attitudes

The students showed enthusiasm when using online learning games to learn, and indicate that learning using online learning games puts them in a better position to understand more about nutrition, and as such, there is need to encourage learning using online learning games. The students, in most cases, do not know much about nutrition, especially when asked about some of the basic elements of nutrition. They, however, indicate their willingness to learn about nutrition using online learning games.

VI. Discussion

This study revealed that teachers encourage the use of technology in teaching nutrition. This is consistent with existing literature. Digital games enhanced their understanding of electromagnetic concepts as compared to students learning similar concepts by traditional methods. It should be mentioned that, nutritional applications could be more effective when guidance is established based on the attitudes, students' achievement, and experience [44]. Additionally, from the findings of this study, it is observed that most of the students have computers and internet access at home. Thus, the internet has made it easy for online learning games' accessibility in many devices, such as computers and smart devices. This could motivate students to learn about nutrition via technology. What these findings demonstrate is that technology has the ability to influence

students' perceptions, as well as the effectiveness of using technology in learning. As a finding of this research, once learning is interesting, students become more motivated. This finding has serious implications for the essential growth and development of students in future, as well as their cognitive development [8].

In addition, games could encourage teachers to be innovative and be more effective. They can identify the best materials through observation and give feedback to the students. The study observation of students playing online learning games is that teachers should teach these lessons in groups because students are more motivated and collaborative when they are playing in groups. There are recommendations for games to hold and present more data and information that helps students to interact directly when playing the game. Moreover, there should be food tips at each game level with rewards. So, the information will be acknowledged by students during playing the game in an efficient manner. Also, nutritional development is promoted through playing interactive games in schools and homes. As a recommendation for future work is to develop an application based on online games to enhance nutritional education then to be included with school curriculum.

VII. Conclusion

This paper was aimed to investigate the attitudes of learners and teachers towards the usage of technology in nutrition education in local primary schools. In addition, it focused on exploring the impact of using online games in order to help develop students awareness of nutritional concepts. The outcome of this paper revealed that the attitude of students and teachers towards using ICT for nutritional learning. With that it has also been discovered that using online learning games is an effective method especially to increase nutritional awareness. Teachers showed willingness into teaching using technology. Most of the participants have computers at home as well as Internet access. Furthermore, The use of online learning games can be an impact and efficient method for advancing knowledge regarding healthy eating habits to students. Game designers have took to interest and advantage of this by accomplishing various games for nutritional learning.

References

[1] Howard-Jones, P., Ott, M., van Leeuwen, T. and De Smedt, B., *The potential relevance of cognitive neuroscience for the development and use of technology enhanced learning*, Learning, media and technology, vol. 40, No, 2, pp. 131–151, 2015.

[2] El-Masri, M. and Tarhini, A., *Factors affecting the adoption of e-learning systems* The Educational Technology Research and Development in Qatar and USA: Extending the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2), Vol. 65, No. 3, pp. 743–763., 2017.

[3] Kirkwood, A. and Price, L., *Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know? A critical literature review*, Learning, media and technology, vol. 39, No. 1, pp. 6–36, 2014.

[4] McKnight, K., O'Malley, K., Ruzic, R., Horsley, M.K., Franey, J.J. and Bassett, K., *Teaching in a digital age: How educators use technology to improve student learning*, Journal of research on technology in education, vol. 48, No. 3, pp. 194–211, 2016.

[5] Beckman, K., Bennett, S. and Lockyer, L., *Understanding students' use and value of technology for learning. Learning*, Media and Technology, vol. 39, No. 3, pp. 346–367, 2014.

[6] DiMaria-Ghalili, R.A., Mirtallo, J.M., Tobin, B.W., Hark, L., Van Horn, L. and Palmer, C.A., *Challenges and opportunities for nutrition education and training in the health care professions: intraprofessional and interprofessional call to action*, The American journal of clinical nutrition, vol. 99, No. 5, pp. 1184S–1193S, 2014.

[7] Vio, F., Fretes, G., Montenegro, E., González, C.G. and Salinas, J., *Prevention of children obesity: a nutrition education intervention model on dietary habits in basic schools in Chile*, Food and Nutrition Sciences, vol. 6, No. 13, p. 1221, 2015.

[8] Benton, D., *The influence of dietary status on the cognitive performance of children Molecular Nutrition Food Research*, vol. 54, pp. 457–470, 2010, Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/mnfr.200900158/epdf.

[9] Adams, K.M., Butsch, W.S. and Kohlmeier, M., *The state of nutrition education at US medical schools*, Journal of Biomedical Education, 2015.

[10] Naghashpour, M., Shakerinejad, G., Lourizadeh, M.R., Hajinajaf, S. and Jarvandi, F. *Nutrition education based on health belief model improves dietary calcium intake among female students of junior high schools*, Journal of health, population, and nutrition, vol. 32, No. 3, p. 420, 2014.

[11] McEvoy, C.S., Cantore, K.M., Denlinger, L.N., Schleich, M.A., Stevens, N.M., Swavely, S.C., Odom, A.A. and Novick, M.B., *Use of medical students in a flipped classroom programme in nutrition education for fourth-grade school students*, Health Education Journal, vol. 75, No. 1, pp. 38–46. 2016.

[12] Rosi, A., Dall'Asta, M., Brighenti, F., Del Rio, D., Volta, E., Baroni, I., Nalin, M., Zelati, M.C., Sanna, A. and Scazzina, F., *The use of new technologies for nutritional education in primary schools: a pilot study*, Public health, vol. 140, pp. 50–55, 2016.

[13] Yang, Y.T.C., Wang, C.J., Tsai, M.F. and Wang, J.S., *Technology-enhanced game based team learning for improving intake of food groups and nutritional elements*, Computers & Education, vol. 88, pp. 143–159, 2015.

[14] Strasburger, V.C., *The new technology revolution: collaborative efforts between pediatricians, schools, and millennials for media education*, In Media Education for a Digital Generation, Routledge, pp. 83–102, 2015.

[15] Savage, A., Februhartanty, J. and Worsley, A., *Adolescent women-a key target population for community nutrition education programs-a qualitative Indonesia case study*, Asia Pacific journal of clinical nutrition, 2016.

[16] Vidgen, H. ed., *Food literacy: key concepts for health and education*, Routledge, 2016.

[17] Hainey, T., Connolly, T.M., Boyle, E.A., Wilson, A. and Razak, A., *A systematic literature review of games-based learning empirical evidence in primary education*, Computers & Education, vol. 102, pp. 202–223, 2016.

[18] Lai, C., *Modelling teachers' influence on learners' self-directed use of technology for language learning outside the classroom*, Computers & Education, vol. 82, pp. 74–83, 2015.

[19] Venkatesh, V., Thong, J.Y. and Xu, X., *Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology*, MIS quarterly, pp. 157–178, 2012.

[20] Thompson, P., *The digital natives as learners: Technology use patterns and approaches to learning*, Computers & Education, vol. 65, pp. 12–33, 2013.

[21] Buckingham, D., *Beyond technology: Children's learning in the age of digital culture*, John Wiley & Sons, 2013.

[22] Mayer, R.E. ed., *The Cambridge handbook of multimedia learning*, Cambridge University press, 2005.

[23] Su, K.D., *An integrated science course designed with information communication technologies to enhance university students' learning performance*, Computers & Education, vol. 51, No. 3, pp. 1365–1374, 2008.

[24] Kelly, A.E., Lesh, R.A. and Baek, J.Y. eds. *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* Routledge. 2014.

[25] Cobern, W.W., *Contextual constructivism: The impact of culture on the learning and teaching of science*, In The practice of constructivism in science education, Routledge, pp. 67– 86, 2012.

[26] Clark, R.E., *Media will never influence learning. Educational technology research and development*, vol. 42, No. 2, pp. 21–29, 1994.

[27] Clark, R.E., *Reconsidering research on learning from media.* Review of Educational Research, vol. 53, No. 4, pp. 445–459, 1983.

[28] Jewitt, C., *Technology, literacy, learning: A multimodal approach*, Routledge, 2012.

[29] Nomass, B.B., *The impact of using technology in teaching English as a second language.* English Language and Literature Studies, vol. 3, No. 1, p. 111, 2013.

[30] Higgins, S., Xiao, Z. and Katsipataki, M. *The impact of digital technology on learning: A summary for the education endowment foundation.* Durham, UK: Education Endowment Foundation and Durham University, 2012.

[31] Hall, R., *Towards a fusion of formal and informal learning environments: The impact of the read/write web.*, Journal of E-learning, vol. 7, No. 1, pp. 29–40, 2009.

[32] Papastergiou, M., *Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation*, Computers & Education, vol. 52, No. 1, pp. 1–12, 2009.

[33] Tamim, R., *Technology integration in the UAE schools: Current status and way forward*, Information Systems Applications in the Arab Education Sector, pp.23-38, 2013.

[34] Archer, K., Savage, R., Sanghera-Sidhu, S., Wood, E., Gottardo, A. and Chen, V., *Examining the effectiveness of technology use in classrooms: A tertiary metaanalysis.* Computers & Education, vol. 78, pp. 140–149, 2014.

[35] Littlejohn, A., Beetham, H. and McGill, L., *Learning at the digital frontier: a review of digital literacies in theory and practice*, Journal of computer assisted learning, vol. 28, No. 6, pp. 547–556, 2012.

[36] Plowman, L., Stevenson, O., McPake, J., Stephen, C. and Adey, C., *Parents, preschoolers and learning with technology at home: some implications for policy*, Journal of computer assisted learning, vol. 27, No. 4, pp. 361–371, 2011.

[37] Hess, S.A., *Digital media and student learning: Impact of electronic books on motivation and achievement*, New England Reading Association Journal, vol. 49, No. 2, p. 35, 2014.

[38] Lysenko, L.V. and Abrami, P.C., *Promoting reading comprehension with the use of technology*, Computers & Education, vol. 75, pp. 162–172. 2014.

[39] Hsu, C.Y., Tsai, C.C. and Wang, H.Y, *Facilitating third graders' acquisition of scientific concepts through digital game-based learning: The effects of self explanation principles*, The Asia-Pacific Education Researcher, vol. 21, No. 1, pp. 71–82, 2012.

[40] Guven, G. and Sulun, Y., *The Effects of Computer-Enhanced Teaching on Academic Achievement in 8th Grade Science and Technology Course and Students' Attitudes towards the Course*, Journal of Turkish Science Education, vol. 9, No. 1, 2012.

[41] Reed, P., Hughes, A. and Phillips, G., *Rapid recovery in sub-optimal readers in Wales through a self-paced computer-based reading programme*, British Journal of Special Education, vol. 40, No. 4, pp. 162–166, 2013.

[42] Tamim, R.M., Bernard, R.M., Borokhovski, E., Abrami, P.C. and Schmid, R.F., *What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study*, Review of Educational research, vol. 81, No. 1, pp. 4–28, 2011.

[43] Contento, I.R., *Nutrition education: linking research, theory, and practice*, Asia Pacific journal of clinical nutrition, vol. 17, No. S1, pp. 176–179, 2008.

[44] Samieipour, S., Rahimzade, R., Naghdi, N., Pakseresht, M., Tavassoli, E., Babaei Heydarabadi, A., Sayehmiri, K., Abedzadeh Zavareh, M.S., Asadi-Samani, M. and Bahmani, M., *Investigation of the effect of education on nutritional knowledge, attitude and performance of primary school students in Ilam- 2015*, Journal of Chemical and Pharmaceutical Sciences, vol. 9, No. 3, pp. 1210–1215, 2016.

[45] Kupolati, M.D., Gericke, G.J. and MacIntyre, U.E.*Teachers' perceptions of school nutrition education's influence on eating behaviours of learners in the Bronkhorstspruit District*, South African Journal of Education, vol. 35, No. 2, pp. 01–10, 2015.

[46] Kupolati, M.D., MacIntyre, U.E. and Gericke, G.J., *School-based nutrition education: features and challenges for success*, Nutrition& Food Science, vol. 44, No. 6, pp. 520–535, 2014.

[47] Rosário, R., Oliveira, B., Araújo, A., Lopes, O., Padrão, P., Moreira, A., Teixeira, V., Barros, R., Pereira, B. and Moreira, P. *The impact of an intervention taught by trained teachers on childhood overweight*, International journal of environmental research and public health, vol. 9, No. 4, pp. 1355–1367, 2012.

[48] Ercan, O., *the effects of multimedia learning material on students' academic achievement and attitudes towards science courses*, Journal of Baltic Science Education, vol. 13, No, 5, 2014.

[49] Cooper, S.B., Bandelow, S. and Nevill, M.E., *Breakfast consumption and cognitive function in adolescent schoolchildren*, Physiology & Behavior, vol. 103, No. 5, pp. 431–439, 2011.

[50] O'Dea, J.A. and Abraham, S.*Knowledge, beliefs, attitudes, and behaviors related to weight control, eating disorders, and body image in Australian trainee home economics and physical education teachers*, Journal of nutrition education, vol. 33, No. 6, pp. 332–340, 2001.

[51] Sharma, P. and Rani, M.U., *Effect of Digital Nutrition Education Intervention on the Nutritional Knowledge Levels of Information Technology Professionals*, Ecology of food and nutrition, vol. 55, No. 5, pp. 442–455. 2016.

[52] Bogden, N,*Technology and Nutrition: Interactive Strategies for Children to Learn Nutrition*, 2015.

[53] Baturay, M.H., Gökçearslan, Ş. and Ke, F.,*The relationship among pre-service teachers' computer competence, attitude towards computer-assisted education, and intention of technology acceptance*, International Journal of Technology Enhanced Learning, vol. 9, No. 1, pp. 1–13, 2017.

[54] Awan, R.N., *What Happens to Teachers ICT Attitudes and Classroom ICT Use when Teachers are made to Play Computer Games?*, International Journal of Information and Education Technology, vol. 1, No. 4, p. 354, 2011.

[55] Mumtaz, S., *Factors affecting teachers' use of information and communications technology: a review of the literature*, Journal of information technology for teacher education, vol. 9, No. 3, pp. 319–342, 2000.

[56] Meerza, A.H. and Beauchamp, G., *Factors influencing attitudes towards information and communication technology (ICT) amongst undergraduates*, An empirical study conducted in Kuwait Higher Education Institutions (KHEIS) 2017.

[57] Rhoda, C. and Gerald, K., *Internal Consistency Reliabilities for 14 computers. Attitude scale*, Journal of Education technology, vol. 14, No. 4, pp. 99–120, 2000.

[58] Bates, A.T., *Technology, e-learning and distance education.*, Routledge, 2005.

[59] Russell, M., Bebell, D., O'Dwyer, L. and O'Connor, K., *Examining teacher technology use: Implications for preservice and inservice teacher preparation*, Journal of teacher Education, vol. 54, No. 4, pp. 297–310, 2003.

[60] Christensen, R., *Effects of technology integration education on the attitudes of teachers and students*, Journal of Research on technology in Education, vol. 34, No. 4, pp. 411–433, 2002.

[61] Balta, N. and Duran, M., *Attitudes of students and teachers towards the use of interactive whiteboards in elementary and secondary school classrooms*, TOJET: The Turkish Online Journal of Educational Technology, vol. 14, No. 2, 2015.

[62] Enayati, T., Modanloo, Y. and Kazemi, F.S.M., *Teachers' attitudes towards the use of technology in education*, Journal of Basic and Applied Scientific Research, vol. 2, No. 11, pp. 10958–10963, 2012.

[63] Dogruer, N., Eyyam, R. and Menevis, I. *The attitudes of English preparatory school instructors towards the use of instructional technology in their classes*, Procedia Social and Behavioral Sciences, vol. 15, No. 2, pp. 5095–5099, 2010.

[64] Ozdamlı, F., Hursen, Ç. and Ozçinar, Z., *Teacher candidates' attitudes towards the instructional technologies*, Procedia-Social and Behavioral Sciences, vol. 1, No. 1, pp.455– 463, 2009.

[65] Kabadayi, A. (2006) *Analyzing Pre-School Student Teachers' and Their Cooperating Teachers' Attitudes towards the Use of Educational Technology*, The Turkish Online Journal of Educational Technology (TOJET), vol. 5, No. 4, 2006.

[66] Zanguyi, S., *Review of teachers' attitudes towards the use of educational technology in teaching process*, Educational Technology, vol. 6, pp. 165–159, 2011.

[67] Seraji, N.E., Ziabari, R.S. and Rokni, S.J.A., *Teacher's Attitudes towards Educational Technology in English Language Institutes*, International Journal of English Linguistics, vol. 7, No. 2, p. 176, 2017.

[68] Chow, P. *Teacher's Attitudes towards Technology in The Classroom*, 2014.

[69] Albirini , *Teachers' attitudes toward information and communication technologies*, The case of Syrian EFL teachers Computers & Education, vol. 47, No. 4, pp. 373–398, 2006.

[70] Buabeng-Andoh, C., *Factors influencing teachers' adoption and integration of information and communication technology into teaching: A review of the literature*, International Journal of Education and Development using Information and Communication Technology, vol. 8, No. 1, p. 136, 2012.

[71] Al-Emran, M., Elsherif, H.M. and Shaalan, K., *Investigating attitudes towards the use of mobile learning in higher education*, Computers in Human Behavior, vol. 56, pp. 93–102, 2016.

[72] Hwang, G.J. and Chang, H.F., *A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students*, Computers & Education, vol. 56, No, 4, pp. 1023–1031, 2011.

[73] Shroff, R.H., Deneen, C.C. and Ng, E.M. *Analysis of the technology acceptance model in examining students' behavioural intention to use an e-portfolio system*, Australasian Journal of Educational Technology, vol. 27, No. 4, 2011.

[74] Tseng, K.H., Chang, C.C., Lou, S.J. and Chen, W.P., *Attitudes towards science, technology, engineering and mathematics (STEM) in a project-based learning (PjBL) environment*, International Journal of Technology and Design Education, vol. 23, No. 1, pp. 87–102, 2013.

[75] Al-Fraihat, D., Joy, M. and Sinclair, J. *Identifying Success Factors for e-Learning in Higher Education*, Proceedings of the 12th International Conference on e-Learning (ICEL), p. 247, 2017.

[76] Condie, R. and Munro, B., *The impact of ICT in schools: Landscape review*, British Educational Communications and Technology Agency (BECTA), Corp Creator, 2007.

[77] Burden, K. and Hopkins, P., *Barriers and Challenges Facing Pre-service Teachers use of Mobile Technologies for Teaching and Learning*, In Blended Learning: Concepts, Methodologies, Tools, and Applications, IGI Global, pp. 1665–1686., 2017.

[78] Osang, F.B., Ngole, J. and Tsuma, C., *February. Prospects and Challenges of Mobile Learning Implementation in Nigeria*, Case Study National Open University of Nigeria NOUN, In International Conference on ICT for Africa, pp. 20–23, 2013.

[79] Blinco, K., Mason, J., McLean, N. and Wilson, S., *Trends and issues in e-learning infrastructure development*, 2004.

[80] Tondeur, J., van Braak, J., Ertmer, P.A. and Ottenbreit-Leftwich, A., *Understanding the relationship between teachers' pedagogical beliefs and technology use in education: a systematic review of qualitative evidence*, Educational Technology Research and Development, vol. 65, No. 3, pp. 555–575, 2017.

[81] Voogt, J. and Knezek, G. eds., *International handbook of information technology in primary and secondary education*, Springer Science & Business Media, vol. 20, 2008.

[82] Sethi, K.K., Bhanodia, P., Mishra, D.K., Badjatya, M. and Gujar, C.P., *Challenges Faced in Deployment of e-Learning Models in India*, In Proceedings of the International Congress on Information and Communication Technology, Springer, Singapore, pp. 647– 655, 2016.

[83] Mayes, R., Natividad, G. and Spector, J.M., *Challenges for educational technologists in the 21st century*, Education Sciences, vol. 5, No. 3, pp. 221–237, 2015.

[84] Anderson, J.L. and Barnett, M. *Learning physics with digital game simulations in middle school science*, Journal of Science Education and Technology, vol. 22, No. 6, pp. 914–926, 2013.

[85] Paraskeva, F., Mysirlaki, S. and Papagianni, A., *Multiplayer online games as educational tools: Facing new challenges in learning*, Computers & Education, 54(2), pp.498-505. 2010.

[86] Campbell, C.,*Study finds smarter students play online games.*, 2016 ,Retrieved from Polygon: https://www.polygon.com/2016/8/8/12406388/online-games-educationbenefit.

[87] Young, M.F., Slota, S., Cutter, A.B., Jalette, G., Mullin, G., Lai, B., Simeoni, Z., Tran, M. and Yukhymenko, M., *Our princess is in another castle: A review of trends in serious gaming for education*, Review of educational research, vol. 82, No. 1, pp.61–89, 2012.

[88] Kapp, K.M., *The gamification of learning and instruction: game-based methods and strategies for training and education*, John Wiley & Sons, 2012.

[89] Squire, K.D., *Video games and education: Designing learning systems for an interactive age*, Educational Technology, pp. 17–26, 2008.

[90] Amory, A., *Game object model version II: a theoretical framework for educational game development*, Educational Technology Research and Development, vol. 55, No. 1, pp. 51–77, 2007.

[91] Kuipers, F., Märtens, M., van der Hoeven, E. and Iosup, A., *The Power of Social Features in Online Gaming*, Social Interactions in Virtual Worlds: An Interdisciplinary Perspective, 2018.

[92] Livingstone, S., Mascheroni, G. and Staksrud, E., *European research on children's internet use: Assessing the past and anticipating the future*, New Media & Society, vol. 20, No. 3, pp. 1103–1122, 2018.

[93] Institute of Digital Media and Child Development Working Group on Games for Health, Baranowski, T., Blumberg, F., Buday, R., DeSmet, A., Fiellin, L.E., Green, C.S., Kato, P.M., Lu, A.S., Maloney, A.E. and Mellecker, R., *Games for health for children—Current status and needed research*, Games for health journal, vol. 5, No, 1, pp. 1– 12. 2016.

[94] Taylor, J., *How technology is changing the way children think and focus.* Psychology Today, 2012, Retrieved from https://www.psychologytoday.com/blog/the-powerprime/201212/how-technology-is-changing-the-way-children-think-and-focus

[95] Granic, I., Lobel, A. and Engels, R.C., *The benefits of playing video games*, American psychologist, vol. 69, No. 1, p. 66, 2014.

[96] KidsHealth, *Mission nutrition*, The Nemours Foundation, 2015, Retrieved from http://kidshealth.org/kid/closet/games/game_nutrition.html.

[97] Annetta, L.A., *Video games in education: Why they should be used and how they are being used*, Theory into practice, vol. 47, No. 3, pp. 229–239, 2008.

[98] Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T. and Boyle, J.M., *A systematic literature review of empirical evidence on computer games and serious games.*, Computers & Education, vol. 59, No. 2, pp. 661–686, 2012.

[99] Whole Kids Foundation, *Awesome eats*, Whole Kids Foundation, 2014, Retrieved from https://www.wholekidsfoundation.org/kids-activities/awesome-eats/

[100] Fisher, C., *Designing games for children: developmental, usability, and design considerations for making games for kids.* 2014, Retrieved from https://books.google.com/books?id=ZxlWBQAAQBAJ&printsec=frontcover&source=gbs_atb#v=onepage&q&f=false

[101] Brannen, J., *Mixing methods: Qualitative and quantitative research*, Routledge. 2017.

[102] Bryman, A., *Quantitative and qualitative research: further reflections on their integration*, In Mixing methods: Qualitative and quantitative research, Routledge, pp. 57–78, 2017.

[103] Choy, L.T. (2014) *The strengths and weaknesses of research methodology: Comparison and complimentary between qualitative and quantitative approaches*, Journal of Humanities and Social Science (IOSR), vol. 19, No. 4, pp. 99–104, 2014.

[104] Kelle, U. and Buchholtz, N., *The combination of qualitative and quantitative research methods in mathematics education: A "mixed methods" study on the development of the professional knowledge of teachers*, In Approaches to qualitative research in mathematics education, Springer, Dordrecht, pp. 321–361., 2015.

[105] McCusker, K. and Gunaydin, S., *Research using qualitative, quantitative or mixed methods and choice based on the research*, Perfusion, vol. 30, No. 7, pp. 537–542, 2015.

[106] Barnham, C., *Quantitative and qualitative research: Perceptual foundations*, International Journal of Market Research, vol. 57, No. 6, pp. 837–854, 2015.

[107] Hammersley, M., *Deconstructing the qualitative-quantitative divide 1 In Mixing methods: Qualitative and quantitative research*, Routledge, pp. 39–55, 2017.

[108] Larson-Hall, J. and Plonsky, L., *Reporting and interpreting quantitative research findings: What get reported and recommendations for the field*, Language Learning, vol. 65, No. S1, pp. 127–159, 2015.

[109] Landrum, B. and Garza, G., *Mending fences: Defining the domains and approaches of quantitative and qualitative research*, Qualitative Psychology, vol. 2, No. 2, p. 199, 2015.

[110] Heale, R. and Twycross, A. *Validity and reliability in quantitative studies*, Evidence-based nursing , 2015.

Gender Differences in the Perception of a Student Information System

Rana Alhajri[1], Ahmed Al-Hunaiyyan[2], Bareeq Alghannam[3], Abdullah Alshaher[4]
Computer Science Department[1]
Computer Science and Information Systems Department[2, 3, 4]
Public Authority for Applied Education and Training, Kuwait

***Abstract*—There is growing recognition that electronic student information systems support college administrations and enhance student performance. These systems must fulfill their user's needs by understanding gender differences among users. This study analyzes gender variations concerning the utilization of online student information systems (SIS), with its central concern being how the dynamics of user experience (UX) are affected. A broad agreement is evident throughout the literature that gender is a crucial aspect when assessing human-computer interactions. Consequently, usability factors are brought into question, although there is some indication among researchers that too much weight is being applied. Study findings are gathered to represent the hedonic and pragmatic qualities of users, with clarifications of students' perspectives deducted from qualitative methods, together with a UX examination made via Kuwait's Public Authority for Applied Education and Training (PAAET) institute. Results suggest that none of the differing approaches and habits the two genders have toward UX should be considered as substantial, with the overall sample recording a perception of UX that is "slightly positive". Furthermore, this research highlights difficulties with usability that developers may wish to take onboard for system upgrades.**

Keywords*—Gender differences; student information system; human-computer interaction; usability; user experience; perceptions

I. Introduction

Systems incorporated into educational facilities need to enhance learning methods by offering those involved an original and dynamic experience filled with a wide range of learning avenues and interests, embracing extra-curricular potential and innovative resources where possible. This needs to be achieved while making the most of e-labs, e-libraries, e-tutoring potentials, simulations, etc. Other beneficial avenues may incorporate Archiving System, as well as include Student Information System (SIS), and e-Advising systems [1], together with learning management systems (LMS) [2]. With SIS, students can access range of functionalities that allow for handling administrative issues – vital for both educators and learners [3, 4]. Studies have confirmed that SIS components make a notable difference in all parties' activities and actions [5, 6].

The key SIS features should be determined, meaning that a fitting evaluation process should be ascertained to make the most of potential. Education facilities should not overlook the importance of integrating with SIS, which has become crucial for robust learning journeys. Its utilization is vital for carrying out a range of college organizational actions, as well as the upgrading of student capabilities. Plus, firm assessment of SIS usability is critical for a wide range of participants, but learners especially, while research carried out within HCI closely relates to the system's ultimate functionality. Assessing how and to what degree a certain system, resource, or service offers usability to those it has been designed for is necessary for both purpose and setting [7]. As a result, developers are required to continue enhancing their systems according to the feedback and assessment of their users, including in cultural and social contexts [8], as well as a personal preference [9], age and gender [10].

A range of varied methods offers guidance in these respects. Still, widely recognized aspects are set out in the literature, for example, with [11], who identifies hedonic and pragmatic qualities as two leading interactive concepts. In this sense, 'pragmatic' concerns task-related aspects and the effectiveness of methods to support particular actions, leading to goal fulfillment. Whereas 'hedonic' concerns those aspects not related to tasks, they are nevertheless crucial for the resource enabling users to pursue their objective. Research covering these dynamics is quite extensive [12, 13, 14], meaning that a range of analyses are available to assist with judging the extent to which systems satisfy their users' goals.

Different characteristics and personalities, for which gender can be important, might affect how learners' approach and utilize online technology. A student's perspective will differ according to their own individual traits, including gender and age, as well as previous experience that dictates the opinions and habits, they develop [15]. Gender, in particular, can prove a defining aspect as far as Kuwait is concerned, as men and women tend to develop different approaches toward technology use. Understanding why this is present in respect of cultural norms can be crucial for aligning systems with a largely conservative society's cultural factors. In general, it can be predicted that both men and women in Kuwait will become accustomed to utilizing online technology to enhance their social fabrics against a backdrop of collectivist culture [16, 17]. However, research into how such mobile-learning behaviors vary between ages and genders has not been extensive enough as far as Kuwait is concerned [18]. This research endeavors to make up for this with an SIS assessment based on gender-driven factors, with a particular focus on UX.

Six sections make up the paper in total. This introduction is followed by a section on research objectives before Section 3 provides a review of the relevant literature. The methodology

is then covered in Section 4, before a discussion of the results in Section 5, and finally a conclusion.

II. Research Objectives

This research has been carried out to clarify learners' perspectives of SIS, together with analyzing their user experience (UX) via the Public Authority for Applied Education and Training (PAAET). It offers the first insight into these dynamics, filling a void in the literature and shining a light on Kuwait's education aspects a result [19]. The focus is on assessing SIS within Kuwait, which has not yet been scrutinized, especially concerning gender factors [18, 20]. Chiefly, we outline two key usability and UX factors: pragmatic and hedonic – i.e., task versus non-task-oriented features [11]. Furthermore, the research explores gender variations concerning opinions of SIS, as well as trialing two vital hypotheses:

H1: Substantial variations exist between the genders when it comes to how students view SIS pragmatically.

H2: Substantial variations exist between the genders when it comes to how students view SIS hedonistically.

This research aims to direct system developers regarding beneficial growth avenues that can further enhance SIS utilization. Such enhancement should extend SIS's efficacy and show how users interact with the various resources, consider both genders, and enrich their understanding of its related capabilities [21].

III. Literature Review

SIS is vital for enabling stakeholders to grasp key details via extensive reports on how both learners and educators are using systems and the various departments involved, including financial aspects. Robust SIS can prove fulfilling for both the educators and students that rely on these systems, as well as having an overall impact on the progression of academic development [22]. The efficacy of the various software and platforms involved is thoroughly examined in related studies [23, 24, 25, 26, 27, 28, 29], with user satisfaction especially driving the research.

The author in [30] carried out a survey at Allama Iqbal Open University (with responses taken from 173 students in total), which explored key aspect of information quality, system quality, service quality, perceived usefulness, and intent use, and user satisfaction. Most respondents were content with the system's functionalities and technical aspects, though they were more critical of the availability of key information and specific system responses [30]. In contrast, [31] assessed SIS's usability links with late student assignments, declined course registrations, and inaccuracies. In doing so, the resulting advice suggested enhancing the system to allow the institution in question to improve these aspects. In addition, research from [32] targeted analyzing the utilization of student information systems via both educators and learners at Yamen's faculty of oil and minerals.

Several studies have indicated differences between the genders when using technology due to cultural habits and beliefs [33, 34, 8]. The author in [35], in New Zealand, looked at comprehending the differing approaches that male and female learners take to a short message service (SMS). The findings showed notable variations in how both genders recognized applicability, together with the purposes of use, although nothing significant in relation to self-efficacy and convenience. Plus, [36] explored the gender dynamics clear from research on smartphone usage carried out across five nations – the USA, Japan, Korea, Italy, and Sweden. Their findings suggest several usage differences according to gender, together with variations in attitude. The author in [37] explores these factors in the context of Saudi Arabia, for which gender variations are apparent because of cultural tendencies. The author in [38] found that users' perspectives are likely to show extensive variations, including age and gender as defining motivations for such preferences. Indeed, it is crucial to appreciate different views caused by age and gender, as well as culture and background [39]. For example, in a study based on Arab GCC nations, [25] examined the driving factors behind female users being drawn to online bulletins to express themselves. Respondents confirmed that corresponding online enabled them to feel more active and stimulated by subjects they might not otherwise wish to raise in a social context. According to the researchers, all these factors were linked to conservative attitudes of broader society.

There is a similar culture to other Arab nations based on a dominant and collectivist approach regarding Kuwait. As proposed by [17], one specific quality of this tendency is the general approach to social situations, family, and friends, which are all ranked highly among personal priorities [17, 40]. The author in [20] examined learners' perspectives on mobile learning potential via 620 responses from HE institutions throughout Kuwait. The results highlight clear variations due to age and gender while revealing related social and cultural factors. Likewise, [16] explored the impact of culture via Instagram concerning gender differences. The findings show that men are more likely to feel comfortable posting confidential material. In contrast, women feel cautious about whether this meets with the values and pressures of a conservative culture.

Quality of use tends to be understood as 'usability.' Trialing this requires concentration on achievement, including the manner of utilization to suit a pragmatic approach [41] relating to the obtainment of behavioral goals [42]. When starting his research two decades ago, [43] considered the trialing of usability methods to overlook the factors of stimulation, user preferences, and innovation. Consequently, he put forward hedonism to add a new layer of understanding, incorporating aspects unrelated to the carrying out of tasks, such as subjective appeal, aesthetics, and novelty [43]. The author in [44] stressed how vital such features might be for overall system appeal. As a result, [45] considered these features to require more focus than pragmatic ones, so they proposed a user experience (UX) model as a suitable long-term analysis tool [45]. The author in [46] also looked at examining how hedonic and pragmatic features impact users, which they did by adopting the UX model to explore how the many variables interconnect and relate.

An alternative familiar approach to analyzing the two competing dynamics is to define them as the goals of usability

versus user experience [47]. A few research analyses were carried out, therefore, according to identified pragmatic and hedonic qualities within both usability and user experience [12, 14, 46, 13]. As proposed by [38], a subjective approach needs to be adopted when considering user experience matters. Those utilizing the technology may develop a wide range of perspectives or go about meeting their objectives in many varied ways. Plus, additional reasons for various perspectives result from both gender and age [12].

IV. Methodology

This section defines the research methodology, setting out the samples, instruments, and procedures that have been employed.

A. Research Sample

In total, 645 respondents contributed to this research, 525 of whom were female and 120 males. These contributors were sourced from the five PAAET colleges: The College of Basic Education, College of Business Studies, College of Technological Studies, College of Health Sciences, and College of Nursing. Due to the colleges in question educating more female than male students in total, the number of female respondents greatly outweighed male ones. For demographic figures and distribution samples, see Table I.

TABLE I. Study Sample (Demographic Data)

		No.	%
Gender	Male	120	18.6
	Female	525	81.4
College	Business Studies	307	47.6
	Health Sciences	89	13.8
	Basic Education	135	20.9
	Technological Studies	79	12.2
	Nursing	35	5.4

B. Research Instruments

This study adopts methods that are both quantitative and qualitative in nature to assess the focus group via a survey. Goal Question Metric (GQM) has been applied to produce the questionnaire statements necessary for surveying the focus group. This is a widely recognized top-down method to examining software metrics via objectives [48], with the objectives being set out according to the defined pragmatic and hedonic usability features. Additionally, to align the questionnaire with answering questions related to gender difference, the User Experience Questionnaire (UEQ) is incorporated [49].

The questionnaire material has been adjusted to suit the context of PAAET students. In doing so, punchy sentences were favored to express the rationale, avoiding only words that might lead to respondents feeling ambiguous toward the intended meaning. The focus research carried out recognized that PAAET students are likely to question material in such a manner. In total, 50 students participated in the preparatory focus group, allowing researchers to ask their own methods and wordings so they could enhance for a larger rollout. This process resulted in the 16 questions edited for the final questionnaire.

There were three sections to the final version, as follows. Section 1 focused on obtaining the demographic data of students, as per their gender and academic institution. Section 2 gathered details to reflect pragmatic behaviors, and Section 3 the hedonistic behaviors. To achieve clarity of response, a five-point Likert-type scale was utilized as follows: 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree. With the questionnaire material edited to best capture the PAAET students' attitudes toward SIS's attitudes, a focus group was then sourced to trial the material's efficacy and upgrade the questions where issues were identified. The questionnaire's applicability was established by defining each theme's interconnections and the representative scores gathered from the 50 participants. SPSS was then also utilized to measure the correlation coefficients. This shows high correlations concerning individual dimensions and the overall score ($p < 0.01$) calculated as between 0.795 to 0.901, which shows high internal consistency and construction integrity.

Likewise, the questionnaire's applicability has been analyzed by identifying Cronbach's alpha via SPSS. Consequently, there are high-reliability levels to the questionnaire, with co-efficient degrees of 0.74–0.93 and an overall Cronbach's alpha reading of 0.96. As a result, the questionnaire material can be considered as fitting the relevant study sample and providing informative results.

C. Research Procedures

With the quantitative questionnaire having been established, a qualitative focus group was put in place. This stage of the process was overseen by a facilitator whose role was to relate the research's purposes and stress how vital the participants' responses were for enhancing SIS. The task then included gaining informed consent, with participants promised that any contributions taken from their responses would be used for nothing else than the defined scientific research. The students were then asked to make introductions, which revealed that some were familiar with each other, which served to enhance the group atmosphere. The subsequent discussion then focused on the questionnaire material they would be asked to respond to. Different volunteers then contributed to reading out a group of statements while the facilitator took notes and identified any queries or issues from the interactions. The participants also contributed with feedback in writing, and the focus group was concluded after around 50 minutes.

The responses and notes that were then gathered were assessed as per the "three coding-framework" of [50]. This enabled the researchers to comprehend many of the current issues and dynamics regarding how systems users are currently approaching their tasks while allowing the questionnaire material to be upgraded to suit. In general, the responses showed that participants found the questions clear and straightforward. Opportunities were still found, however, to remove or to blend some of the statements, meaning that the material was enhanced as a result. With the focus group process complete, the questionnaire could then be administered online with approval from PAAET's higher administration. With this approval obtained, all faculties then received the

questionnaire with directions for rolling out to their students. A seven-day response period followed before the feedback was examined via SPSS, including frequency, percentage, mean, standard deviation (SD), and t-test.

V. Results and Discussion

The questionnaire results are detailed here, focusing on those responses that capture the SIS perspectives. Furthermore, the research hypothesis is discussed concerning gender variations.

A. Students' SIS Perceptions

The findings from our assessment of participants' SIS perceptions are presented in this section. To provide a basis for analyzing the responses, the two categories of pragmatic and hedonic are applied. The sub-sections below contain tables to show percentages, means, standard deviations (SD), t-test, and how each item ranks in dimension due to the average mean values. Taken together, the data highlights variations between the genders in respect of their responses.

1) Pragmatic quality: Table II contains statistics relating to SIS's pragmatic features (task-oriented features), together with the capabilities relating to achieving 'do-goals' and its applicability regarding a range of possible tasks [11]. Analyzing the mean values identified within the Table II data (items 1 to 8) shows that participants tended to record a neutral-to-positive perspective of SIS and its worth, which did not seem to be affected by gender. In contrast, a neutral-to-disagree perspective was recorded for Question 1 'All system commands are executed quickly,' with 2.79 being the average mean. The highest rank was achieved by system security, with participants showing 4.01 as an average mean value (as per item 8). Ease of use (item 5) came second, recording an average mean of 3.6; and system accuracy came third, recording an average mean of 3.44 (as per Question 6). Also, feedback for whether participants felt they had adequate SIS training (Question 4) produced a neutral-to-agree response and an average mean of 3.18.

Taking the overall average mean of 3.36 from the pragmatic results, the research found that participants were slightly satisfied with the functions offered by SIS, together with its efficacy and usability. Based on this, Hypothesis 1 was analyzed, but Table II data highlights no substantial variations between the genders with one exception. For 'The SIS is an easy-to-use program' (item 5), the findings show significant gender variations for 'level of significance (p-0.03) in favor of female (t-test) $p < 0.05$.

2) Hedonic quality: Table III contains the findings relating to the hedonic aspects of SIS (the non-task-related UX features), which shows the system's recognized capabilities in terms of aiding its users' objective– in this example, the system's aesthetics [11]. Assessing the mean values indicated within the table shows that participants have a neutral-to-positive perspective of SIS attractiveness, together with innovation, with no clear variation between the genders. The perspectives recorded on how data is graphically represented by SIS (Question 13), achieved the highest average mean value (4.16), with excitement coming second (Question 9, with an average mean of 3.29). The third place was taken by asking how interesting the system is (Question 10, with an average mean of 3.20). In contrast, the question regarding creativity only came seventh (Question 16, with an average mean of 3.13). In the last place was the question on attractiveness (Question 11, with an average mean of 2.95).

TABLE II. SIS Evaluation "Pragmatic Quality"

	Question	Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Mean	SD	Sig.	Rank
Q1	All system commands are executed quickly.	male	10	16	33	42	19	2.86	1.380	0.52	8
		female	58	67	143	190	67	2.77	1.428		
Q2	I believe that the SIS meets my requirements.	male	16	18	22	43	21	3.29	1.292	0.82	5
		female	71	78	96	202	78	3.26	1.266		
Q3	I think the SIS is practical and effective.	male	17	18	24	37	24	3.28	1.328	0.87	6
		female	73	76	102	193	81	3.25	1.275		
Q4	I got enough training on how to use the SIS.	male	19	21	17	44	19	3.19	1.337	0.95	7
		female	68	101	92	195	69	3.18	1.256		
Q5	The SIS is an easy-to-use program.	male	14	17	25	35	29	3.40	1.312	0.03	2
		female	29	66	94	202	134	3.66	1.149		
Q6	The SIS performs my registration accurately.	male	12	17	23	44	24	3.43	1.241	0.81	3
		female	38	75	111	213	88	3.45	1.143		
Q7	The SIS is reliable.	male	14	20	21	37	28	3.38	1.322	0.86	4
		female	57	69	108	190	101	3.40	1.242		
Q8	The SIS is secured.	male	5	8	16	46	45	3.98	1.077	0.68	1
		female	25	8	74	240	178	4.02	0.985		

TABLE III. SIS EVALUATION "HEDONIC QUALITY"

	Question	Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Mean	SD	Sig.	Rank
Q9	The screen of SIS is exciting.	male	10	16	33	42	19	3.37	1.152	0.41	2
		female	58	67	143	190	67	3.27	1.171		
Q10	The SIS is an interesting system.	male	13	15	28	42	22	3.38	1.230	0.10	3
		female	82	80	106	186	71	3.16	1.284		
Q11	The SIS interface is attractive.	male	14	34	34	26	12	2.90	1.170	0.65	8
		female	82	108	146	129	60	2.96	1.239		
Q12	The SIS is stimulating.	male	14	22	33	31	20	3.18	1.248	0.49	5
		female	53	86	130	184	72	3.26	1.184		
Q13	Graphics showing students' performance is challenging.	male	6	5	12	43	54	4.12	1.078	0.62	1
		female	11	26	58	200	230	4.17	0.954		
Q14	The SIS is an interesting system	male	17	20	30	34	19	3.15	1.281	0.47	6
		female	66	85	117	170	87	3.24	1.263		
Q15	The SIS is an innovative system.	male	16	21	34	28	21	3.14	1.279	0.13	4
		female	41	100	124	165	95	3.33	1.198		
Q16	The SIS is a creative system.	male	18	24	24	32	22	3.13	1.341	0.95	7
		female	60	109	129	159	68	3.13	1.214		

Calculating the overall average mean for SIS's hedonistic qualities gives us 3.31, recording a neutral-to-slightly satisfying response regarding innovation, attractiveness, and stimulation. Taking a t-test to look more closely at gender dynamics (based on Table III data) highlights no substantial variation between genders regarding 'level of significance.' Plus, a t-test result of $p<0.05$ establishes that the results do not confirm Hypothesis 2.

B. Pragmatic vs. Hedonic Qualities of the SIS

The two categories of pragmatic and hedonic are utilized for analyzing the interactive qualities of SIS. In this context, the pragmatic tends to be linked to task-oriented features. In contrast, hedonic indicates features unrelated to the task's users carry out, but which nevertheless prove crucial to the attractiveness and interactivity offered in pursuit of such objectives [11]. The findings detailed in the sections above highlight respondents' perspectives regarding SIS in respect of both pragmatic and hedonic contexts. Their feedback shows that SIS is considered essential in respect of enabling learners to register courses and access their records to suit deadlines and the completion of quality work. Comparing the scores shows that feedback on pragmatic aspects results in a slightly higher average mean of 3.36 against 3.31 – as per Fig. 1.

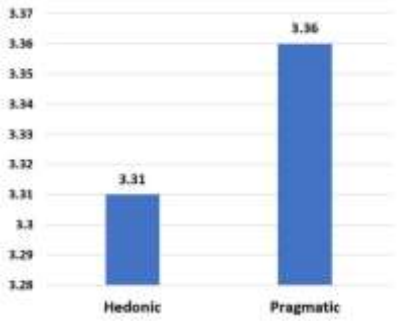


Fig. 1. Pragmatic vs. Hedonic Quality of the SIS.

The findings highlight the need for SIS enhancements to meet users' expectations in a manner that upgrades efficacy, usability, attractiveness, innovation, productivity, and learnability. Productivity relates to speed and convenience, enabling learners to complete their objectives swiftly. In contrast, efficacy concerns the precision they are allowed in meeting their goals without being held back by data entry issues or performance restrictions [51]. Numerous methods allow developers to attempt such SIS upgrades. To achieve ease-of-use within a system, those working on its features need to allow students the means of adapting to new features without encountering major learning curves along the way. The most effective route to securing such upgrades is to focus on SIS aspects that sync with the user's current capabilities. Indeed, developers should provide staged upgrades to functionality so that the technology never lacks familiarity. The researchers of [52, 53] all advise educational institutions to consider effective training and guidance where possible so that learners can always make the most of their systems' potentials. Therefore, the development and release of training advice via online videos and tutorials and more focused training programs for both genders should enable seamless SIS upgrades.

The results also offer guidance for prioritizing SIS aesthetics by introducing dynamic and attractive new functionalities. Including creativity within developers' enhancements can prove to define when it comes to engaging with students, handling issues in a way that is both innovative and enticing. The research of [54] also highlights creativity as integral to upgrading the efficacy and usability of SIS, highlighting numerous positives resulting from presenting software and innovations in a versatile and artistic manner. The authors in [55] and [56] both focus on attractiveness as essential despite being a hedonistic quality, identifying numerous benefits in how students comprehend and utilize the tools and materials available. Furthermore, the efforts taken to achieve quality graphics help to stimulate users due to appealing visuals. Any additions that allow for speedier

comprehension of instructions or data help users take charge of the tools at hand [57]. According to Human-Computer Interaction research, the benefits of quality aesthetics on learners' subjective impressions and their subsequent responses are widely acknowledged. Extensive research points to the potential of aesthetic interfaces heightening engagement levels that users can achieve [58, 59, 60]. As per [59], a practical approach to aesthetics is advised because users respond to such dynamics.

C. Differences According to Gender

This examination records Hypothesis 1 as being reached in part, finding no substantial variation between the genders, apart from concerning participants' usability – as explored in the section regarding pragmatic qualities that resulted in substantial variations with a 'level of significance' ($p=0.03$), in favor of female (t-test) $p < 0.05$. For Hypothesis 2, however, no substantial variations were identified regarding how either gender views SIS's hedonistic qualities.

The variations that do exist between genders can affect how they approach and utilize online technology. Learners' perspectives differ according to various user features, for example, individual characteristics, cognitive tendencies, age and gender, and previous experience that can shape opinions of and behaviors toward SIS [15]. Furthermore, the utilization of online tools may vary between the genders because of societal dynamics. For example, in Kuwait, it is predicted that both men and women will utilize online media for social purposes before any other use as a result of belonging to a collectivist culture, which is also affected by the country's education system being gender-segregated. Despite this research identifying SIS evaluation results that can be drawn upon to help designers upgrade systems, the findings also point to the genders showing no favoritism in terms of either aesthetics or functionality.

VI. Conclusion

According to SIS perspectives, this research has examined gender variations – an essential component to any modern educational facility – via learners' opinions and responses to its various dynamics. The UX positives and negatives recorded according to SIS's use within PAAET institutions have been analyzed under the context of pragmatic or hedonic use, which stand for the two vital elements of an effective system. By combining qualitative and quantitative approaches, the responses gathered from 645 PAAET participants informed the results. Also, for qualitative purposes, a focus group meeting took place to gain some prior insight into participants' utilization and opinions of the applicable SIS to refine the materials that would make up the questionnaire. When combined, the focus group findings and the survey data statistical analysis suggest that the students' opinions toward SIS were slightly positive.

Regarding UX dynamics, the pragmatic qualities are considered slightly more favorable than hedonic ones – 3.36 versus 3.31. The findings call for SIS upgrades to meet users' objectives, with a particular focus on creativity and attractiveness among the innovative steps taken. Consequently, throughout the PAAET facilities, the SIS is no longer completely applicable to the learning objectives and delivery methods that will bring the best out of learners. More effective and engaging systems need to be developed so that students can realize their academic potential, specifically via attractive visual dashboards and many other features enhanced by better quality graphics. As a result, the constant review and upgrading of SIS features via wide-scale research and analyses are crucial if systems remain robust. Continuous feature improvements regularly further the productive nature of students' SIS use via interactivity.

This research's findings are applicable to usability developers' concerns, together with any professionals with a vested interest in how SIS use varies between the genders. Two key hypotheses have been trialed in a gender context to represent both pragmatic and hedonic SIS features. Notable variations were identified between the genders as far as pragmatic features are concerned, but not with hedonic features. The results heighten awareness of SIS potential within PAAET, particularly regarding developers' requirements to carry out UX assessments. The example included which was shown to be applicable and effective in informing upgrades based on context. Upgrading and innovating systems based on creativity and attractiveness will make SIS more accessible and engaging to both genders, having a knock-on effect on educational performance via value and productivity.

Acknowledgment

This research was supported and funded by the Public Authority for Applied Education and Training, project number: BS-19-03.

References

[1] A. Al-Hunaiyyan, A. Bimba and S. Al-Sharhan, "A cognitive knowledge-based model for an adaptive e-advising system.," Interdisciplinary Journal of Information, Knowledge, and Management (IJIKM), Volume 15, pp. 247-263, 2020.

[2] A. Al-Hunaiyyan, S. Al-Sharhan and R. Al-Hajri, "Prospects and Challenges of Learning Management Systems in Higher Education," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No. 12, pp. 73-79, 2020.

[3] S. Mukerjee, "Student information systems – implementation challenges and the road ahead," Journal of Higher Education Policy and Management, 34(1), 51–60. https://doi.org/10.1080/1360080X.2012.642332, p. 51–60, 2012.

[4] S. Rochimah, H. Rahmani and U. Yuhana, "Usability characteristic evaluation on administration module of Academic Information System using ISO/IEC 9126 quality model 2015," in International Seminar on Intelligent Technology and Its Applications (ISITIA). doi: 10.1109/ISITIA.2015.7220007, Surabaya, 2015.

[5] C. Guarin, E. Guzman and F. Gonzalez, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining," IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje, 10 (3), pp. 119-125, 2015.

[6] H. Widodo , M. Kertahadi and I. Suyadi, "The Influence of Job Relevant Information, Task Technology Fit, and Ease of Use Information Technology Due to the User Performance: A Case Study on the and Use of Academic and Financial Information System in University of Brawijaya," Asian Journal of Social Sciences & Humanities, 4 (2) , pp. 128-138, 2015.

[7] J. Nielsen, Designing User Interfaces for International Use, New Yory\k: Elsevier, 1990.

[8] R. Alhajri, S. Al-Sharhan, A. Al-Hunaiyyan and T. Alothman, "Design of educational multimedia interfaces: individual differences of learners,"

in Proceedings of the Second Kuwait Conference on e-Services and e-Systems, Kuwait, 2011.

[9] N. Al-Huwail, S. Al-Sharhan and A. Al-Hunaiyyan, "Learning Design for a Successful Blended E-learning Environment: Cultural Dimensions," INFOCOMP. Journal of Computer Science, Volume 6 – No. 4, pp. 60-69, 2007.

[10] A. Al-hunaiyyan, S. Al-Sharhan and R. Alhajri, "Instructors Age and Gender Differences in the Acceptance of Mobile Learning," International Journal of Interactive Mobile Technologies (iJIM). Vol. 11, No. 4, 2017.

[11] M. Hassenzahl, "User experience (UX): Towards an experiential perspective on product quality," in Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine on - IHM '08 DOI:10.1145/1512714.1512717., New York, 2008.

[12] A. Hinderks, M. Schrepp and J. Thomaschewski, "A Benchmark for the Short Version of the User Experience Questionnaire," in Proceedings of the 14th Inernational Conference on Web Information (WEBIST), 2018.

[13] M. Rauschenberger, M. Schrepp, M. Cota, S. Olschner and J. Thomaschewski, "Efficient Measurement of the User Experience of Interactive Products. How to use the User Experience Questionnaire (UEQ). Example: Spanish Language Version," International Journal of Interactive Multimedia and Artificial Intelligence. Vol. 2, Nº 1. DOI: 10.9781/ijimai.2013.215, pp. 39-45, 2013.

[14] M. Schrepp, A. Hinderks and J. Thomaschewski, "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S)," International Journal of Interactive Multimedia and Artificial Intelligence. 4.103, 2017.

[15] R. Alhajri, A. Alhunaiyyan and E. AlMousa, "Understanding the Impact of Individual Differences on Learner Performance Using Hypermedia Systems," International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 12(1). doi:10.4018/IJWLTT.2017010101, pp. 1-18, 2017.

[16] A. Al-Kandari, A. Al-Hunaiyyan and R. Alhajri, "The Influence of Culture on Instagram Use," Journal of Advances in Information Technology, vol. 7, no. 1, pp. 54-57, 2016.

[17] A. Al-Kandari, F. Al-Sumait and A. Al-Hunaiyyan, "Ali A. Al-Kandari, Fahad Y. Al-Sumait & Ahmed Al-Hunaiyyan (2017) Looking perfect: Instagram use in a Kuwaiti cultural context," Journal of International and Intercultural Communication, Volume 10, Issue 4, https://doi.org/10.1080/17513057.2017.1281430, pp. 273-290, 2017.

[18] F. Dashti and A. Aldashti, "EFL College Students' Attitudes towards Mobile Learning," International Education Studies, vol. 8, no. 8, pp. 13-20, 2015.

[19] S. Al-Sharhan, A. Al-Hunaiyyan and H. Al-Sharrah, "A new efficient blended e-learning model and framework for k12 and higher education: Design and," in 2010 fifth international conference, 2010.

[20] A. Al-Hunaiyyan, S. Al-Sharhan and R. Alhajri, "Instructors Age and Gender Differences in the Acceptance of Mobile Learning," International Journal of Interactive Mobile Technologies (iJIM). Vol. 11, No. 4, 2017.

[21] P. Morville, "User Experience Design," 2014. [Online]. Available: http://semanticstudios.com/user_experience_design/.

[22] D. Demirkol and C. Seneler, "Evaluation of a Student Information System (SIS) in terms of User Emotions, Performance and Perceived Usability: A Pilot Study," in XV. European Conference on Social and Behavioral Sciences, Kusadasi, Turkey, 2018.

[23] M. Gemmell and R. Pagano:, "A Post-Implementation Evaluation of a Student Information System in the UK Higher Education Sector," The Electronical Journal of Information Systems Evaluation, 6(2), 2003.

[24] M. R. Nordaliela, H. Suriani and E. L. Nathaniel, "Usability Analysis of Students Information System in a Public University.," Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 4(6), pp. 806-810, 2013.

[25] I. Sherifi, "Impact of information systems in satisfying students of the university: Case study from Epoka University," European Journal of Business and Social Sciences, pp. 167-175, 2015.

[26] A. Alzahrani, I. Mahmmud,, T. Ramayah, O. Alfarraj and N. Alalwan, "Alzahrani, A., Mahmmud, I., Ramayah, T., Alfarraj, O., & Alalwan, N. (2017). Modelling digital library success using the DeLone and McLean information system success model.," Journal of Librarianship and Information Science 51(2), 2017.

[27] C. Gurkut and M. Cemal Nat, "Important Factors Affecting Student Information System Quality and Satisfaction," Eurasia Journal of Mathematics, Science and Technology Education.14(3), pp. 923-932, 2017.

[28] S. Tabrizi, C. Tufekci, O. Gumus and A. Cavus, "Usability Evaluation for Near East University Student Information System. , pp 235-243," New Trends and Issues Proceedings on Humanities and Social Sciences. 03, pp. 235-243, 2017.

[29] D. Demirkola and C. Seneler, "Evaluation of Student Information System (SIS) In Terms of User Emotion, Performance and Perceived Usability: A Turkish University Case (An Empirical Study)," Procedia Computer Science 158, pp. 1033-1052, 2019.

[30] K. Mir and A. Mehmood, ". (2016). Examining the Success Factors of Online Student Support System at AIOU.," in Pan-Commonwealth Forum 8 (PCF8), KLCC, KL. Malaysia, 2016.

[31] A. Eludire, "The Design and Implementation of Student Academic Record Management System," Research Journal of Applied Sciences, Engineering and Technology, Osun State, Nigeria, vol. 3, no. 8, pp. 707-712, 2011.

[32] A. Farid, "Improve the usability of student information system at Aden Universit," International Journal of Contemporary Computer Research (IJCCR), Vol.1 Issue.1, 2016.

[33] H. Hijazi-Omari and R. Ribak, "PLAYING WITH FIRE: On the domestication of the mobile phone among Palestinian teenage girls in Israel," Information, Communication & Society Vol. 11 , Issue 2, 2008.

[34] N. Baron and Y. Hård af Segerstad, "Cross-cultural patterns in mobile phone use: Public space and reachability in Sweden, the USA, and Japan," New Media & Society 12(1), pp. 13-34, 2010.

[35] T. Goh, "Exploring Gender Differences in SMS-Based Mobile Library Search System Adoption," Educational Technology & Society, 14 (4), p. 192–206, 2011.

[36] N. Baron and E. Campbell, " Gender and mobile phones in cross-national context," Language Sciences 34 (2012), p. 13–27, 2012.

[37] E. W. Baker, S. S. Al-Gahtani and G. S. Hubona, "The effects of gender and age on new technology implementation in a developing country: Testing the theory of planned behavior (TPB)," Information Technology & People, vol. 20, no. 4, p. 352–375, 2007.

[38] G. Boy, The Hand book o fHuman-Machine Interaction: A Human-Centered Design Approach, 1st ed. edition, Milton: CRC Press, 2017.

[39] Prayaq, "The Importance of User Experience Design," 2019. [Online]. Available: https://uxplanet.org/the-importance-of-user-experience-design-988faf6ddca2?gi=59cd019477c9.

[40] A. Al-Hunaiyyan, Design of Multimedia Software in Relation to Users' Culture. Ph.D thesis, University of Hertfordshire, UK, 2000.

[41] J. Lewis and J. Sauro, "what's the difference between pragmatic and hedonic usability?," 17 May 2020. [Online]. Available: https://measuringu.com/pragmatic-hedonic/.

[42] P. Zimmermann, Beyond Usability – Measuring Aspects of User Experience. Thesis for: PhD, Zurich: Swiss Federal Institute of Technology Zurich, 2008.

[43] M. Hassenzahl, A. Platz, M. Burmester and K. Lehner, "Hassenzahl, Marc & Platz, Axel & Burmester, Michael & Lehner, Katrin. (2000). Hedonic and ergonomic quality aspect determine a software's appeal," in Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems. The Netherlands, April 1-6, 2000. 10.1145/332040.332432., 2000.

[44] M. Hassenzahl, "The Effect of Perceived Hedonic Quality on Product Appealingness," International Journal of Human–Computer Interaction, 13:4, DOI: 10.1207/S15327590IJHC1304_07, pp. 481-499, 2001.

[45] S. Kujala, V. Roto, K. Väänänen-Vainio-Mattila and A. Sinnelä, "2001. Identifying Hedonic Factors in Long-Term User Experience," in Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces. June 22–25, 2011. https://doi.org/10.1145/2347504.2347523, Milan, Italy, 2011.

[46] T. Merčun and M. Žumer, "Exploring the influences on pragmatic and hedonic aspects of user experience," in Proceedings of the Ninth International Conference on Conceptions of Library and Information Science, June 27-29, 2016, Uppsala, Sweden, 2016.

[47] J. Preece, Y. Rogers and H. Sharp, Interaction Design: Beyond Human-Computer Interaction, 4th Edition, Indiannapolis: Wiley & Sons, Inc, 2015.

[48] V. Basili, G. Caldiera and D. Rombach, "The Goal Question Metric Approach," Encyclopedia of software engineering, pp. 528-532, 1994.

[49] B. Laugwitz, T. Held and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire," in HCI and Usability for Education and Work. USAB 2008. Lecture Notes in Computer Science, vol 5298, Berlin, Heidelberg, Springer, 2008.

[50] T. Nyumba, K. Wilson, C. Derrick and N. Mukherjee, "The use of focus group discussion methodology: Insights from two decades of application in conservation," Methods in Ecology and Evolution (MEE), 9, p. 20–32, 2018.

[51] W. Quesenbery, "What Does Usability Mean: Looking Beyond 'Ease of Use'," 12 December 2020. [Online]. Available: https://www.wqusability.com/articles/more-than-ease-of-use.html.

[52] S. Al-Sharhan, A. Al-Hunaiyyan, R. Alhajri and N. Al-Huwail, "Utilization of Learning Management System (LMS) Among Instructors and Students," in Advances in Electronics Engineering. Lecture Notes in Electrical Engineering, vol 619, Singapore, Springer, 2020, pp. 15-23.

[53] A. Johnson, M. Jacovina, D. Russell and C. Soto, "Challenges and solutions when using technologies in the classroom," in Adaptive educational technologies for literacy instruction, New York, Taylor & Francis, 2016, pp. 13-29.

[54] R. McDaniel, R. Fanfarelli and R. Lindgren, "Creative Content Management: Importance, Novelty, and Affect as Design Heuristics for Learning Management Systems," IEEE Transactions on Professional Communication, vol. 60, no. 2, pp. 183-200, 2017.

[55] M. Baharum, Z. Zainul Rashid, Z. Husin, S. Sahat and Z. Abu, "An evaluation of Universiti Teknologi Mara branch campuses websites towards acceptance among staff," in in 2011 IEEE Conference on Open Systems, 2011.

[56] Y. He, C. Cheng, Q. Xu and L. Yang, "A research on methods and applications of case study in public administration," in International Conference on Management Science Engineering 21th Annual Conference Proceedings, Helsinki, Finland, 2014.

[57] MindTools.com, "Charts and Graphs: choosing the Right Visual For Your Data," 2020. [Online]. Available: https://www.mindtools.com/pages/article/Charts_and_Diagrams.htm. [Accessed 15 3 2020].

[58] C. Miller, "Aesthetics and E-Assessment: the interplay of emotional design and learner performance," Distance Education ;32, p. 307–337, 2011.

[59] M. Thielsch, R. Haines and L. Flacke, "Experimental investigation on the effects of website aesthetics on user performance in different virtual tasks," PeerJ, 7, e6516, 2019.

[60] P. Van Schaik and J. Ling, "Modelling user experience with web sites: usability, hedonic value, beauty and goodness," Interacting with Computers, 2008;20, p. 419–432, 2008.

Student Information System: Investigating User Experience (UX)

Ahmed Al-Hunaiyyan[1], Rana Alhajri[2], Bareeq Alghannam[3], Abdullah Al-Shaher[4]
Computer Science and Information Systems Department, College of Business studies[1, 3, 4]
Computer Science Department, Higher Institute for Telecommunication and Navigation[2]
Public Authority for Applied Education and Training, Kuwait

***Abstract*—There is growing recognition that electronic student information systems support college administrations and enhance student performance. These systems must fulfill their user's needs (efficiently achieve their academic goals) while also providing a positive user experience (UX). This study used quantitative and qualitative approaches to elucidate students' perceptions and investigate UX toward the SIS currently used at the Public Authority for Applied Education and Training (PAAET), a higher education institution in Kuwait. Survey data collected from 645 PAAET students were analyzed to determine their perceptions of and experiences using this SIS. The findings revealed that students had a slightly positive UX with this SIS. The system's perspicuity, stimulation, and dependability were rated slightly higher than its novelty, attractiveness, and efficiency. The most pertinent usability issues that focus on the human interaction with systems were identified and discussed, hoping that it will allow officials and SIS system developers alike to make relevant and impactful improvements to newer versions of these systems. These results shed light on the need for continuous SIS evaluation and a broad research scope to develop innovative SIS with intelligent functions for novel activities. Such features enhance students' interactivity and productivity, which encourage their academic success.**

***Keywords*—Student information system; user experience; usability; human-computer interaction; e-learning**

I. Introduction

Successful student information systems (SIS) make students productive and improve the workflow of their academic services [1]. These systems, including learning management systems (LMS), provide functions and tools that overcome college-level administrative and academic problems [2, 3]. SIS allows college students to manage their data, including registering in courses, maintaining grades, showing transcripts, and generating progress reports. Although SIS are widely used in the academic world, these systems require regular evaluation to make them more productive. Having effective and efficient SIS significantly impacts stakeholder groups' operation and performance [4, 5]. Therefore, the key features of SIS must be identified, and appropriate evaluation criteria must be developed to measure them.

Usability is associated with the user acceptability of any system [6]. Determining the usability aspects is essential because millions of people, including students, instructors, and administrative staff, use SIS to conduct administrative and academic tasks. Recently, user experience (UX) has gained considerable attention among researchers in academia and industry and has become a vital aspect of the products' success [7]. The author in [8] stated that UX is considered a key aspect in designing products and services. It is argued that institutions that apply UX design activities in their system development achieve many potential advantages that increase user satisfaction. The author in [9] believes that an effective UX does not appear on its own but must be systematically evaluated. Due to its importance, several frameworks and models have been proposed to design and assess the UX of interactive systems. These models serve as a guide to improve the design and the quality of interactive systems [10].

Although the literature provides UX evaluations with various information systems, it does not do so for SISs [11]. Designing usable SIS is essential; however, little research was conducted, especially in universities among Arab Gulf countries. Several usability studies did not analyze and develop such systems considering students' perceptions and their UX. This observation led to the work presented in this paper, which tries to fill this gap by investigating student experience with SIS. This study was conducted to elucidate students' perceptions of UX with the SIS used at the Public Authority for Applied Education and Training (PAAET). It is a pioneer study given the absence of research on this topic and the context of Kuwait's educational system [12]. Its significance is to provide system developers with pertinent improvement possibilities for future versions of this SIS to enhance efficiency and attractiveness and improve users' interactions with the system and its related functions [13].

This article is organized into sections. Section 2 reviews the relevant literature; Section 3 explains the methodology. The results and a discussion thereof are presented in Section 4, and Section 5 draws conclusions and explores future directions.

II. Literature Review

A. Evaluation of SIS

One of the critical systems for managing HE's administrative and academic aspects the SIS [14]. Although SIS are widely used in the academic world, these systems require constant evaluation to ensure their relevance and effectiveness [15, 11]. An effective SIS not only satisfies administrators and students but also ensures sustainable academic progress [1]. Determining the usability level of an SIS from the human-computer interaction perspective is an essential consideration for universities. Developers, therefore, need to continually be creating better, more usable systems

informed by understanding their potential users with concerns of social and cultural issues [16]; individual differences [17]; and gender and age differences [18].

Some research focuses on SIS development, while others investigate SIS usability, UX, and perceptions. The author in [19] described the design and development of a novel SIS. The study was motivated by the fact that there are difficulties associated with the manual methods used to manage student information at the University of Diyala and aimed at adopting new SIS to increase efficiency and accuracy that also helps college administrations speed up the decision-making process. Besides, [20] developed an SIS for the Faculty of Electronics & Computer Engineering, University Teknikal Malaysia Melaka. They described the development steps needed to operate the system. Their system focused on recording and updating students' records system replacing the old traditional SIS. They believed that this system would contribute to new knowledge in the field, ease use, and better arrangement and scheduling.

Considering the usability of SIS, [21] conducted a study to examine the usability factors of an SIS at a public university. Data were collected from 132 computer science students using a questionnaire. The authors used factor analysis, which involves the user's perceptions of usefulness, speed, interface, and error corrections. The results demonstrate that several usability attributes, such as the importance of information and system functionalities that are commonly gathered, affect user engagement. A similar study was carried out by [22] to examine an SIS's usability at Near East University. The results provide recommendations to improve the interface and enhance system attractiveness. In addition, an empirical study conducted by [11] elucidated the influence of student background on SIS experiences in terms of emotion, performance, and perceived usability. Substantial variations between user emotion, performance, and perceived usability were found.

The author in [14] states that SIS is critical when managing the administrative and academic aspects of HE. Their study investigated how system quality, information quality, and information presentation impact academic and administrative staff satisfaction. Data collected from 120 users were evaluated using factor analysis and regression, revealing that system quality and information quality have significant indirect effects on user satisfaction while information delivery does not directly or indirectly. The author in [23] evaluated SIS performance toward improving the current SIS productivity at Kalinga State University using an interview-based approach to elucidate students, administrators, and instructors' perceptions. The system was found to satisfy five usability factors: usefulness, functionality, reusability, maintainability, and security.

B. User Experience (UX)

Usability and UX are often confused. The author in [24] believes that usability is mainly concerned with the functional part of a system. At the same time, UX is related to how the users interact with a system that involves the user's feelings and attitudes. Similarly, [13] claimed that UX focuses on understanding users, their needs, interests, strengths, and limitations. She stresses that investigating UX helps to improve users' interactivity with the system and raises their perceptions. Similarly, Norman and Nielsen stated that UX involves the users' perception of usability, which examines how users view the usefulness and effectiveness of the system or application [8, 25, 26].

The author in [7] developed a User Experience Questionnaire (UEQ) that measures UX. The six scales of this questionnaire comprehensively represent UX by quantifying six dimensions of usability. These dimensions are attractiveness "the product should look attractive, enjoyable, friendly, and pleasant"; efficiency "the user should perform tasks with the product fast, efficient and in a pragmatic way"; perspicuity "the product should be easy to understand, clear, simple, and easy to learn"; dependability "the interaction with the product should be predictable, secure and meets my expectations"; stimulation "using the product should be interesting, exciting and motivating"; novelty "the product should be innovative, inventive and creatively designed" [27]. The author in [28] investigated the impact of culture on the UX of a system using the UEQ. They examined how Indonesian and German students evaluate common systems according to their UX and provided insights and possible explanations for any detected cultural differences. Other studies that have used the UEQ to evaluate UX include [7, 29, 30].

III. Methodology

The research methodology is described in this section, including the research sample, instruments, and procedure.

A. Research Sample

This study included 645 participants from the five PAAET colleges: College of Basic Education, College of Business Studies, College of Technological Studies, College of Health Sciences, and College of Nursing. Table I presents the demographic data and sample distribution of the study population (gender and college).

TABLE I. Study Sample's Demographics (N = 645)

Characteristic	Categories	n	%
Gender	Male	120	18.6
	Female	525	81.4
College	Business Studies	307	47.6
	Health Sciences	89	13.8
	Basic Education	135	20.9
	Technological Studies	79	12.2
	Nursing	35	5.4

B. Research Instruments

This study used both quantitative and qualitative approaches to assess the UX of PAAET's SIS. A focus group was conducted before administering the online questionnaire to the entire study population.

1) Focus group: A focus group was administered to gain confidence in a tentative UX questionnaire that is to be used as a tool to evaluate the UX of the SIS. A single face-to-face focus group session was administered by a facilitator at the

College of Business Studies (CBS), one of PAAET's colleges, during the fall term of the 2019-2020 academic year. Sixteen CBS students with over 30 credits were chosen randomly to participate in the focus group session to validate the questionnaire's statements. The questionnaire's statements were discussed and verified in terms of content for validity by the 16 participants. Statements that seemed ambiguous or redundant were highlighted. The resulting in-depth discussion within the focus group provided further insight into the students' perceptions of the on-line system in question. The findings were documented in a report approved by the participants five days after the focus group discussion.

2) Questionnaire: The questionnaire used for this study was adapted from the UEQ designed to measure UX that can be found at www.ueq-online.org. The author in [7] reported that the UEQ reliably depicts six UX dimensions, including attractiveness, efficiency, perspicuity, dependability, stimulation, and novelty. For this study, the authors chose to use the six dimensions differently than described in the handbook while still following its recommendations to "use terms that fit the language of your stakeholders" [7]. The items were modified to reflect the specific nature of PAAET students. The rationale was explained in short sentences, rather than singular words, to avoid students questioning the meaning of the phrase, which is the expected behavior of PAAET students, confirmed by the focus group. Fifty students piloted the adapted UX questionnaire during the fall semester of the 2019-2020 academic year while the focus group was administered. The objective was to find any ambiguity in the statements and alter them accordingly. The final version of the questionnaire, consisting of 23 statements mapped onto six scales, was shaped by the focus group's findings and the pilot study.

The final questionnaire used in this study consists of seven parts. Part 1 collects students' demographic information (gender and college). Parts 2 to 7 evaluate the UX of the SIS system and assess the six usability dimensions rated using a five-point Likert-type scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree). The questionnaire was developed to measure students' perceptions of PAAET's SIS. A pilot study was conducted to test the validity and reliability of the questionnaire. The questionnaire's internal consistency was confirmed by determining the correlations between each theme and the questionnaire's total score, obtained from surveying 50 students. The researchers used SPSS to calculate the correlation coefficients in Table II. The correlations between the individual dimensions and the overall score were high ($p < 0.01$) and ranged from 0.795 to 0.901, indicating high internal reliability and construction integrity.

Similarly, the questionnaire's reliability was established by calculating the Cronbach's alpha for each dimension using SPSS (Table III). The questionnaire's dimensions exhibited a high degree of reliability, with coefficients that ranged from 0.74 to 0.93. The total Cronbach's alpha score was 0.96; therefore, the questionnaire was reliable and generalizable.

3) Research procedures: The researchers developed the quantitative UX questionnaire and conducted the qualitative focus group discussion. The focus group session was organized and run by a facilitator who began by stating the study's objective and emphasized the students' feedback on improving the SIS. Informed consent was obtained, and the students were assured that the information extracted from their feedback would be used for scientific research only. The facilitator instructed the participants to introduce themselves, some of whom knew each other, which seemed to improve the group dynamics. The focus group discussion was guided by the sequence of the statements in the UEQ. Each group of statements within a construct was carefully read out loud by different participants. The data collection relied on the facilitator's notetaking during team discussions and the participants' written comments. The session lasted for about 50 minutes.

The focus group feedback was analyzed using the "three coding-framework" reported by [31] . This helped to understand the current state, problems, and opportunities for the system and help shape the questionnaire's final statements. The focus group students simultaneously found the statements in the questionnaire clear, easy to follow, and doable. Nevertheless, they all have agreed on the need to merge/delete some statements, as shown in Table IV.

The focus group session helped to shape the final version of the questionnaire. The questionnaire was administered online during the fall term of the 2019-2020 academic year. After obtaining approval from PAAET's higher administration, the questionnaire was distributed to all faculty, who, in turn, were instructed to forward it to their students. Responses were collected over seven days. The results were analyzed using SPSS and frequency, percentage, mean, standard deviation (SD) was used in the analysis.

TABLE II. CORRELATION BETWEEN UX DIMENSION

Dimension	Correlation
Attractiveness	0.837**
Efficiency	0.839**
Perspicuity	0.795**
Dependability	0.892**
Stimulation	0.901**
Novelty	0.867**

**p < 0.01

TABLE III. CRONBACH'S ALPHA OF EACH UX DIMENSION

Dimension	No. of Items	Cronbach's Alpha
Attractiveness	4	0.90
Efficiency	5	0.86
Perspicuity	6	0.74
Dependability	10	0.85
Stimulation	5	0.82
Novelty	4	0.93
Total Score	34	0.96

TABLE IV. CHANGES MADE TO THE QUESTIONNAIRE

Dimension	Action
Attractiveness	Merge statements 1, 2, and 3 and use only statement 1
Efficiency	No changes
Perspicuity	Merge statements 10 and 11 in one statement
Dependability	No changes
Stimulation	Delete statement 19
Novelty	Delete statement 21

IV. RESULTS AND DISCUSSIONS

A. Students' Perceptions of the SIS

This section presents the results of the analysis of students' perceptions of the SIS. Tables V to X include 23 items distributed in the six dimensions: attractiveness; efficiency; perspicuity; dependability; stimulation; and novelty. The tables presented in the following sub-sections show the percentages, means, standard deviations (SD), and ranks of the items within the dimensions according to their mean values.

1) Attractiveness: Attractiveness refers to whether the system looks appealing and pleasant to the user. Table V lists the three items used to investigate the attractiveness dimension. The mean values of items A1 were the highest, which indicated that the system screen was exciting (mean = 3.29). Item A2 that the SIS was interesting, came in second (mean = 3.20), while A3 "The SIS interface is attractive" scored the lowest in this dimension (mean = 2.95). Visual design is a non-functional element designing interfaces and confers attractiveness to any given system [32, 33]. The analysis of students' responses summarized in Table V, revealed that the SIS' attractiveness was marginally appreciated. The attractiveness dimension was ranked fifth of the six dimensions as the mean value was 3.14, slightly higher than the neutral point 3.0.

Aesthetics is a set of principles that relate to a design's attractiveness. The visual design includes consistency, color, association, pattern, scale, outline, and visual weight. It engages users by helping them to perform the correct functionality on the system smoothly [34]. System designers should use aesthetics to enhance their designs' usability, innovation, and attractiveness [35, 33]. Visual design is a crucial success factor; however, its importance has changed over time. The author in [36] investigate the dynamics between the significance and the attractiveness dimensions of software product features and their influence on user satisfaction. The study provided useful insight into the trade-offs between the attractiveness and importance dimensions and informs which features should be focused on evolving software products.

2) Efficiency: Three questions were used to examine the efficiency of the SIS (Table VI). Efficiency implies that users can perform their tasks quickly and without unnecessary effort. As for students' responses, item E2 "I believe that the SIS meets my requirements" was ranked highest with a mean value of 3.27, which indicates that the SIS is marginally efficient and meets students' needs. Item E3 also showed that the SIS was ranked slightly effective (mean = 3.26). However, item E1 "All system commands are executed quickly" was ranked lowest with a mean value of 2.78, which was below students' expectations. Efficiency is essential to usability, which measures how quickly users can accomplish their tasks and, as such, positively impacts system quality [37]. The efficiency dimension was ranked lowest with an overall mean value of 3.10. This indicates that students are neutral to agree that the SIS is efficient slightly.

3) Perspicuity: Perspicuity refers to the simplicity of the system, easy to use, and easy to learn. Table VII shows five items used to investigate perspicuity. Item P2 "It is necessary to have a clear explanation of how to use the SIS" ranked highest with a mean value of 4.07. In contrast, item P1 indicated that students were neutral concerning whether they received enough training to use the SIS and ranked the lowest (mean = 3.18). Perspicuity considers how easy it is for users to learn to perform a task using the interface and how easy it is to remember how to perform it. This dimension was ranked the highest of the dimensions with an overall mean value of 3.58, which indicated that students moderately agree that the system is easy to use and learn.

Learnability is measured by the level of ease with which users become proficient at using a system [38]. The author in [39] stated that learnability is one of the five quality dimensions of usability; the others are efficiency, memorability, satisfaction, and error. Students moderately agreed that the SIS was easy to use, systems commands were understood, and the system can be used without the help of others. During the focus group session, a few students suggested conducting training and orientation sessions on using the system's functions. Others said that they did not use all the system functions, focusing on basic functions that allow registrations and viewing their grades and schedules. According to [40], the learnability and user-friendliness of a system are inversely proportional to the amount of training time needed for its use. Focusing on the design helps to increase learnability and ease of use by allowing users to understand the interface quickly without training. Besides, consistency in interface design makes the system's menus and commands well organized and easy to use; inconsistencies can confuse systems.

Training and guidance are critical issues for the proper use of technology in educational institutions. It is stressed by [41, 42], that colleges and universities should provide adequate training and guidance for students and advisors to use and utilize the systems' tools and functions.

4) Dependability: Dependability refers to whether the user feels in control of the system and the interaction with the system is predictable. Table VIII includes four items used to investigate dependability, namely, system reliability, expectancy, accuracy, and security. The students' responses to item D4, "The SIS is secured" (mean = 4.02) were the highest. In contrast, item D3 "The SIS meets my expectations," ranked lowest with a mean value of 3.17. The dependability dimension ranked third of the six dimensions with a mean

value of 3.51. This indicated that students moderately agree that the system is trustworthy.

Dependability is a non-functional property of a system derived mainly from whether users can trust the system. An alternative concept that also contributes to dependability is avoiding system failures that are more frequent and severe than acceptable [43]. Dependability encompasses many attributes, such as the system's reliability, availability, durability, accuracy, and security [44]. Software designers should value this dimension highly because dependable software is often praised and recommended by its users. The author in [45] claims that dependability can provide services that can be trusted defensibly within a given timeframe.

5) Stimulation: Stimulation queries whether the system is exciting, motivating, and fun to use. The four items in Table IX show that the SIS was moderately stimulating. Item S2, "Displaying the number of courses that I have completed, and the remainder is valuable," ranked first (mean = 4.16). Second, item S3's mean score of 3.57 indicated that the SIS motivated students to do better. Item S4, "The SIS is an interesting system," ranked lowest with a mean value of 3.22. The stimulation dimension was ranked second of the six dimensions with a mean value of 3.55. This indicated that students moderately agreed that the system was stimulating. To achieve a successful design that has a positive impact on the user and achieves a business objective, persuasive elements must be explicitly considered in the context of the behavior that the application seeks to influence; and this must take place in the early stages of the design process. The author in [46] listed motivational drivers for system developers, which are an excellent place to start with any application: "collecting, connecting, achievement, feedback, reciprocity, and blissful productivity."

6) Novelty: Novelty reflects whether a system is innovative and creative. Four items are shown in Table X investigating novelty. From the mean values of each item, the students' responses to item N3 "The SIS is technically advanced" ranked highest with a mean value of 3.51. This indicated that students slightly agree that the system is novel. Novelty can catch the user's attention and is defined by [26] as "The quality of being new, original, or unusual." Software novelty can help a system be noticed among the many other systems and applications; however, to do so, the system must also be useful for users. Reference added other aspects of novelty that contribute to UX, such as creation, invention, and innovation.

TABLE V. STUDENTS' PERCEPTIONS OF THE SIS'S "ATTRACTIVENESS"

		Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree				Rank
No.	Questions	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Mean	SD	
A1	The screen of SIS is exciting.	68	10.5	83	12.9	176	27.3	232	36.0	86	13.3	3.29	1.168	1
A2	SIS is an interesting system.	95	14.7	95	14.7	134	20.8	228	35.3	93	14.4	3.20	1.276	2
A3	The SIS interface is attractive.	96	14.9	142	22.0	180	27.9	155	24.0	72	11.2	2.95	1.225	3

TABLE VI. STUDENTS' PERCEPTIONS OF THE SIS'S "EFFICIENCY"

	Questions	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree		Mean	SD	Rank
No.		Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%			
E1	All system commands are executed quickly.	172	26.7	133	20.6	87	13.5	169	26.2	84	13.0	2.78	1.418	3
E2	I believe that the SIS meets my requirements.	87	13.5	96	14.9	118	18.3	245	38.0	99	15.3	3.27	1.270	1
E3	I think the SIS is practical and effective.	90	14.0	94	14.6	126	19.5	230	35.7	105	16.3	3.26	1.284	2

TABLE VII. STUDENTS' PERCEPTIONS OF THE SIS'S "PERSPICUITY"

	Questions	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree		Mean	SD	Rank
No.		Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%			
P1	I got enough training on how to use the SIS.	87	13.5	122	18.9	109	16.9	239	37.1	88	13.6	3.18	1.270	5
P2	It is necessary to have a clear explanation of how to use the SIS.	32	5.0	31	4.8	62	9.6	257	39.8	263	40.8	4.07	1.068	1
P3	The SIS can be used, and its contents understood without the help of others.	52	8.1	114	17.7	132	20.5	205	31.8	142	22.0	3.42	1.234	4
P4	The commands and links on the SIS are clear and understandable.	31	4.8	75	11.6	135	20.9	275	42.6	129	20.0	3.61	1.077	3
P5	The SIS is an easy-to-use program.	43	6.7	83	12.9	119	18.4	237	36.7	163	25.3	3.61	1.184	2

TABLE VIII. STUDENTS' PERCEPTIONS OF THE SIS'S "DEPENDABILITY"

No.	Questions	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree		Mean	SD	Rank
		Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%			
D1	The SIS performs my registration accurately.	50	7.8	92	14.3	134	20.8	257	39.8	112	17.4	3.45	1.161	2
D2	The SIS is reliable.	71	11.0	89	13.8	129	20.0	227	35.2	129	20.0	3.39	1.256	3
D3	SIS meets my expectations.	84	13.0	119	18.4	144	22.3	200	31.0	98	15.2	3.17	1.264	4
D4	The SIS is secured.	30	4.7	16	2.5	90	14.0	286	44.3	223	34.6	4.02	1.002	1

TABLE IX. STUDENTS' PERCEPTIONS OF THE SIS'S "STIMULATION"

No.	Questions	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree		Mean	SD	Rank
		Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%			
S1	The SIS is stimulating.	67	10.4	108	16.7	163	25.3	215	33.3	92	14.3	3.24	1.196	3
S2	Displaying the number of courses that I have completed, and the remainder is valuable.	17	2.6	31	4.8	70	10.9	243	37.7	284	44.0	4.16	0.977	1
S3	The SIS motivated me to perform better in my courses.	50	7.8	69	10.7	151	23.4	216	33.5	159	24.7	3.57	1.192	2
S4	The SIS is an interesting system	83	12.9	105	16.3	147	22.8	204	31.6	106	16.4	3.22	1.266	4

TABLE X. STUDENTS' PERCEPTIONS OF THE SIS'S "NOVELTY"

No.	Questions	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree		Mean	SD	Rank
		Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%			
N1	The SIS is an innovative system.	57	8.8	121	18.8	158	24.5	193	29.9	116	18.0	3.29	1.215	2
N2	SIS is a creative system.	78	12.1	133	20.6	153	23.7	191	29.6	90	14.0	3.13	1.237	4
N3	The SIS is technically advanced.	63	9.8	71	11.0	123	19.1	251	38.9	137	21.2	3.51	1.218	1
N4	SIS is an innovative program.	75	11.6	113	17.5	153	23.7	202	31.3	102	15.8	3.22	1.241	3

During the focus group session, some students expressed that while the SIS allowed them to achieve their intended goal, the interface looks traditional and boring. Some said the SIS does not provide intelligent "what if" scenarios concerning course requirements and scheduling or suggestions to boost their GPAs. An intelligent expert can provide personalized support to each student and creates a virtual collaborative environment that includes advisors, students, registrars, and IT staff to ensure that the SIS effectively contributes to student success. It is essential to inject creativity into the design of these systems. The author in [47] stresses that creative design can improve the efficiency and utilization of a system and claim that there are benefits to using creative approaches to design and develop innovative new models for software presentation and information retrieval.

B. Comparison of UX Dimensions

The results of the analysis revealed that the students had a marginally positive perception of the SIS. A comparison of the six dimensions of UX is illustrated in Fig. 1. The mean values are used to indicate the level of the six dimensions of UX: perspicuity, mean 3.58; stimulation, mean 3.55; dependability, mean 3.51; novelty, mean 3.29; attractiveness, mean 3.14; and efficiency, mean 3.10.

For the six dimensions, the means ranged from 3.10 to 3.58, with an overall average of 3.36. Three dimensions were above 3.5, which suggests that the SIS was slightly appreciated. With the ongoing evolution of tools and applications, software improvements are essential, and research must best inform the most pressing. Emphasis should be placed on the efficiency, attractiveness, and novelty dimensions of this SIS.

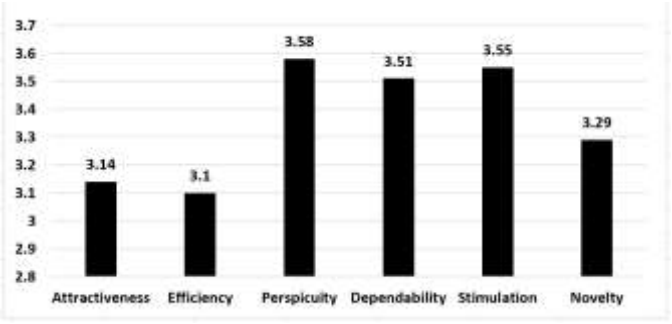


Fig. 1. Mean Values of the UX Dimensions.

V. Conclusions and Future Directions

The current study investigated the UX of an SIS, a key platform in any contemporary academic institution's operation, by analyzing students' perceptions. The strengths and weaknesses of the design/usability/UX provided by the SIS currently used at PAAET were examined according to six factors that are central to successful systems.

Both qualitative and quantitative methods were used to query a sample of 645 students from the five PAAET colleges. For the former, focus group discussions were used to explore students' experiences and perceptions of the SIS and improve the questionnaire used to collect data on six crucial aspects of UX for the latter. Taken together, the results from the focus group and statistical analysis of the survey data indicate that the participants had a positive impression of the SIS. As for UX dimensions, the findings revealed that this SIS's perspicuity, stimulation, and dependability were rated slightly higher than its novelty, attractiveness, and efficiency. This suggests that the SIS used at PAAET since 2010 can no longer support the new learning models and delivery modes that students require. Students need efficient systems to collect their academic data in easy-to-use visual dashboards with attractive features such as graphical data representation. These results shed light on the need for continuous SIS evaluation and a broad research scope to develop innovative SIS with intelligent functions for novel activities. Such features enhance students' interactivity and productivity, which encourage their academic success.

This study's results are interest to usability experts and those studying user behavior and practical uses of interactive systems. A poorly designed user interface, ineffective mobile experience, and lack of service availability can turn the SIS into a source of frustration. However, an effective UX earns users' interest and, most importantly, enhances their productivity. The present study provides insight into PAAET's SIS specifically and, in general, highlights the need for educational institutions to perform regular SIS UX evaluations, such as the one illustrated here, which proved to be a valid and reliable way of generating context-specific recommendations.

Future work should focus on designing and implementing intelligent SIS. Intelligent services using adaptive, knowledge-based feedback creates a personalized experience that accommodates individual needs. Innovative SIS also provides what-if scenario analysis, tracks student data trends, provides students with insight, and demonstrates academic progress. Also, the mobile experience with SIS must be reliable, as this interface is only gaining relevance, and some students had negative experiences when using PAAET's SIS on their mobile device. Moreover, redesigning the system according to new creative and innovative approaches and using a more attractive layout injected with stimulating elements that render the interface more user-friendly would enhance the SIS's look and feel while also improving its efficiency and efficacy.

Acknowledgment

This research was supported and funded by the Public Authority for Applied Education and Training, project number: BS-19-03.

References

[1] Demirkol and C. Seneler, "Evaluation of a Student Information System (SIS) in terms of User Emotions, Performance and Perceived Usability: A Pilot Study," in XV. European Conference on Social and Behavioral Sciences, Kusadasi, Turkey, 2018.

[2] S. Rochimah, H. Rahmani and U. Yuhana, "Usability characteristic evaluation on administration module of Academic Information System using ISO/IEC 9126 Quality Model 2015," in International Seminar on Intelligent Technology and Its Applications (ISITIA), doi: 10.1109/ISITIA.2015.7220007, Surabaya, 2015.

[3] S. Al-Sharhan, A. Al-Hunaiyyan, R. Alhajri and N. Al-Huwail, "Utilization of Learning Management System (LMS) Among Instructors and Students," in Advances in Electronics Engineering. Lecture Notes in Electrical Engineering, vol 619, Singapore, Springer, 2020, pp. 15-23.

[4] C. Guarin, E. Guzman and F. Gonzalez, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining," IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje, 10(3), pp. 119-125, 2015.

[5] H. Widodo, M. Kertahadi and I. Suyadi, "The Influence of Job Relevant Information, Task Technology Fit, and Ease of Use Information Technology Due to User Performance: A Case Study on the Use of Academic and Financial Information Systems at the University of Brawijaya," Asian Journal of Social Sciences & Humanities, 4(2), pp. 128-138, 2015.

[6] J. Nielsen, Designing User Interfaces for International Use, New York: Elsevier, 1990.

[7] A. Hinderks, M. Schrepp, F. Mayoa, M. Escalona and J. Thomaschewski, "Developing a UX KPI based on the user experience questionnaire," Computer Standards & Interfaces, 65, pp. 38-44, 2019.

[8] D. Norman and J. Nielsen, "The Definition of User Experience (UX)," 2020. [Online]. Available: https://www.nngroup.com/articles/definition-user-experience/.

[9] B. Alenljung, J. Lindblom, R. Cort and T. Ziemke, "User Experience in Social Human-Robot Interaction.," International Journal of Ambient Computing and Intelligence, 8, pp. 12-31, 2017.

[10] S. Tasoudis and M. Perry, "Participatory Prototyping to Inform the Development of a Remote UX Design System in the Automotive Domain," Multimodal Technologies Interact, 2(4), p. ADD PAGE NUMBERS, 2018.

[11] D. Demirkol and C. Seneler, "Evaluation of Student Information System (SIS) In Terms of User Emotion, Performance and Perceived Usability: A Turkish University Case (An Empirical Study)," Procedia Computer Science 158, pp. 1033-1052, 2019.

[12] S. Al-Sharhan, A. Al-Hunaiyyan and H. Al-Sharrah, "A new efficient blended e-learning model and framework for k12 and higher education: Design and," in 2010 fifth international conference, 2010.

[13] P. Morville, "User Experience Design," 2014. [Online]. Available: http://semanticstudios.com/user_experience_design/.

[14] C. Gurkut and M. Cemal Nat, "Important Factors Affecting Student Information System Quality and Satisfaction," Eurasia Journal of Mathematics, Science and Technology Education, 14(3), pp. 923-932, 2017.

[15] K. Mir and A. Mehmood, "Examining the Success Factors of Online Student Support System at AIOU," in Pan-Commonwealth Forum 8 (PCF8), KLCC, KL. Malaysia, 2016.

[16] R. Alhajri, S. Al-Sharhan, A. Al-Hunaiyyan and T. Alothman, "Design of educational multimedia interfaces: individual differences of learners," in Proceedings of the Second Kuwait Conference on e-Services and e-Systems, Kuwait, 2011.

[17] N. Al-Huwail, S. Al-Sharhan and A. Al-Hunaiyyan, "Learning Design for a Successful Blended E-learning Environment: Cultural Dimensions," INFOCOMP. Journal of Computer Science, Volume 6 – No. 4, pp. 60-69, 2007.

[18] A. Al-hunaiyyan, S. Al-Sharhan and R. Alhajri, "Instructors Age and Gender Differences in the Acceptance of Mobile Learning," International Journal of Interactive Mobile Technologies (iJIM). Vol. 11, No. 4, 2017.

[19] I. Hassan, "Design and Implement a Novel Student Information Management System – Case Study," International Journal of Computer Science and Mobile Computing, 7(7), pp. 20-31, 2018.

[20] N. Hashim and S. Mohamed, "Development of Student Information System," International Journal of Science and Research (IJSR), 2(8), pp. 256-260, 2013.

[21] M. R. Nordaliela, H. Suriani and E. L. Nathaniel, "Usability Analysis of Students Information System in a Public University.," Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS), 4(6), pp. 806-810, 2013.

[22] S. Tabrizi, C. Tufekci, O. Gumus and A. Cavus, "Usability Evaluation for Near East University Student Information System," New Trends and Issues Proceedings on Humanities and Social Sciences, 3, pp. 235-243, 2017.

[23] Bayangan-Cosidon, "Student Information System for Kalinga State University-Rizal Campus," International Journal of Management and Commerce Innovations, 4(1), pp. 330-335, 2016.

[24] R. Berezhnoi, "Differences Between Usability and User Experience," 2019. [Online]. Available: https://f5-studio.com/articles/difference-between-usability-and-user-experience/.

[25] Prayaq, "The Importance of User Experience Design," 2019. [Online]. Available: https://uxplanet.org/the-importance-of-user-experience-design-988faf6ddca2?gi=59cd019477c9.

[26] D. Qualls, "Novelty and Innovation: UX Design for Long-Term Results," 2015. [Online]. Available: https://medium.com/@DanoQualls/novelty-and-innovation-ux-design-for-long-term-results-46412c7e9de5.

[27] A. Hinderks, M. Schrepp and J. Thomaschewski, "User Experience Questionnaire (UEQ)," [Online]. Available: https://www.ueq-online.org/.

[28] H. Santoso, M. Schrepp, A. Hinderks and J. Thomaschewski, "Cultural Differences in the Perception of User Experience," in Conference: Mensch und Computer, Regensburg, Germany, 2017.

[29] A. Hinderks, M. Schrepp and J. Thomaschewski, "A Benchmark for the Short Version of the User Experience Questionnaire," in Proceedings of the 14th International Conference on Web Information (WEBIST), 2018.

[30] M. Schrepp, A. Hinderks and J. Thomaschewski, "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S)," International Journal of Interactive Multimedia and Artificial Intelligence, 4, p. 103, 2017.

[31] T. Nyumba, K. Wilson, C. Derrick and N. Mukherjee, "The use of focus group discussion methodology: Insights from two decades of application in conservation," Methods in Ecology and Evolution (MEE), 9, p. 20–32, 2018.

[32] N. Ngadiman, S. Sulaiman and W. Wan Kadir, "A systematic literature review on attractiveness and learnability factors in web applications," in IEEE Conference on Open Systems (ICOS), doi: 10.1109/ICOS.2015.7377272, Bandar Melaka, 2015.

[33] A. Al-Hunaiyyan, Design of Multimedia Software in Relation to Users' Culture. Ph.D thesis, University of Hertfordshire, UK, 2000.

[34] R. Alhajri and A. Al-Hunaiyyan, "Integrating Learning Style in the Design of Educational Interfaces," ACSIJ Advances in Computer Science: an International Journal, Vol. 5, Issue 1, No.19 , January 2016. ISSN : 2322-5157, 2016.

[35] F. Fagerholm, A. Hellas, M. Luukkainen, K. Kyllonen, S. Yaman and H. Mäenpää, "Designing and implementing an environment for software start-up education: Patterns and anti-patterns.," Journal of Systems and Software, 146, pp. 1-13, 2018.

[36] A. Kumar, "Software product features: Should we focus on the attractive or the important?," Journal of Decision Systems, 24(4), DOI: 10.1080/12460125.2015.1080587, pp. 449-469, 2015.

[37] A. Kaur, P. Grover and A. Dixit, "Performance Efficiency Assessment for Software Systems," in Software Engineering. Advances in Intelligent Systems and Computing, 731, Singapore, Springer, 2019.

[38] R. Harrison, D. Flood and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," J Interact Sci, 1(1), doi: 10.1186/2194-0827-1-1, p. ADD PAGE NUMBERS, 2013.

[39] A. Joyce, "How to Measure Learnability of a User Interface," 2019. [Online]. Available: https://www.nngroup.com/articles/measure-learnability/.

[40] V. Batchu, "Learnability in UX and how it makes wonders with the users," 2019. [Online]. Available: https://uxdesign.cc/learnability-in-ux-and-how-it-makes-wonders-with-the-users-95833c8bf951.

[41] A. Johnson, M. Jacovina, D. Russell and C. Soto, "Challenges and solutions when using technologies in the classroom," in Adaptive educational technologies for literacy instruction, New York, Taylor & Francis, 2016, pp. 13-29.

[42] A. Al-Hunaiyyan, S. Al-Sharhan and R. Al-Hajri, "Prospects and Challenges of Learning Management Systems in Higher Education," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No. 12, http://dx.doi.org/10.14569/IJACSA.2020.0111209, pp. 73-79, 2020.

[43] H. Alkaraawi, "Solution of Dependability of Computer Systems in Bases of Computer Science," International Journal of Engineering and Management Sciences (IJEMS), 8(2), pp. 140-147, 2017.

[44] S. Bernardi, J. Merseguer and D. Petriu, "Dependability Modeling and Analysis of Software Systems Specified with UML," ACM Computing Surveys, 45(1), doi: 10.1145/2379776.2379778, p. ADD PAGE NUMBERS, 2012.

[45] Z. Avizienis, J. Laprie, B. Randell and C. Landwehr, "Basic Concepts and Taxonomy of Dependable and Secure Computing," IEEE Transactions on Dependable and Secure Computing, 1, pp. 11-33, 2004.

[46] G. Zichermann, "Gamification at Work: Designing Engaging Business Software," 2020. [Online]. Available: https://www.interaction-design.org/literature/book/gamification-at-work-designing-engaging-business-software/chapter-5-58-motivation.

[47] R. McDaniel, R. Fanfarelli and R. Lindgren, "Creative Content Management: Importance, Novelty, and Affect as Design Heuristics for Learning Management Systems," IEEE Transactions on Professional Communication, 60(2), pp. 183-200, 2017.

Mitigating Denial of Service Signaling Threats in 5G Mobile Networks

Raja Ettiane[1], Rachid EL Kouch[2]
National Institute of Posts and
Telecommunication

***Abstract*—With the advent of 5th generation (5G) technology, the mobile paradigm witnesses a tremendous evolution involving the development of a plethora of new applications and services. This enormous technological growth is accompanied with an huge signaling overhead among 5G network elements, especially with emergence of massive devices connectivity. This heavy signaling load will certainly be associated with an important security threats landscape, including denial of service (DoS) attacks against the 5G control plane. In this paper, we analyse the performance of a defense mechanism based randomization technique designed to mitigate the impact of DoS signaling attack in 5G system. Based on massive machine-type communications (mMTC) traffic pattern, the simulation results show that the proposed randomization mechanism decreases significantly the signaling data volume raised from the new 5G Radio Resource Control (RRC) model under normal and malicious operating conditions, which up to 70% while avoiding the unnecessary resource consumption.**

***Keywords*—5G New Radio (NR) network; Radio Resource Control (RRC) state model; Denial of Service (DoS); signaling threats; randomization**

I. Introduction

The emergence of the 5G standard was accompanied with a phenomenal rise in traffic volumes emanating from various new services and applications. To meet these new challenges, 5G technology has introduced three new classes of services, namely, the enhanced mobile broadband (eMBB), massive machine-type communications (mMTC) and ultra-reliable low latency communications (URLLC) [1]. While the eMBB services will ensure an enhanced throughput, the mMTC services will handle massive number of connected devices with stringent energy efficiency and battery autonomy constraints, and the URLLC use case will provide low latency and high reliability services [2]. These new 5G challenging requirements will certainly increase the complexity of the management procedures designed to handle the rising demand of mobile subscribers.

To reduce network signaling complexity and unnecessary control transmissions, ongoing research works are progressing in many fronts with the aim of optimizing the signaling load for a robust and ultra-lean 5G designs. Indeed, a novel radio resource control (RRC) inactive state $RRC_{INACTIVE}$ have been introduced for Next Generation of Radio Access Network (NG-RAN) [3] to enhance the energy efficiency, reduce the latency and optimize the signaling load through optimizing the idle-to-connected state transition. Even if the new 5G RRC model was developed to meet the huge signaling overhead handled by the cellular paradigm, the short inactivity timers joined to the tremendous number of connected devices will entail a number of security flaws, including the problem of Denial of Service (DoS) attacks against the next generation of radio access network (NG-RAN) signaling control plane, named signaling threats. The DoS signaling threats were first emerged in 3G system [4], [5], [6], [7], involving the signaling attack that exploits the Radio Access Bearer (RAB) allocation/ release procedures to overload 3G entities, specifically the Radio Network Controller (RNC) entity. By using the well known network parameter, named inactivity timer $T_{5G_{inac}}$, this attack could be also carried out against the 5G system to overload the signaling control plane, which can disturb the network functionality giving rise to a productivity loss for network operator.

Several research works have tackled the problem of signaling threats in 3G/4G mobile networks and have proposed detection and defense mechanisms to mitigate the impact of such attacks [4], [8], [10], but little research efforts have been dedicated to signaling-based threats in 5G context. A survey of the 5G security architecture related to the primary protocols of the control plane signalling was presented in [11], [12]. In [13], the authors have proposed a defence mechanism to protect the paging protocols against security and privacy attacks [14]. The proposed solution aims at securing the 4G/5G devices from unauthorized/fake paging messages by introducing a new identifier, named P-TMSI, randomizing the paging occasions, and conceiving a symmetric-key based broadcast authentication framework. In [15], the issue of DoS signaling attacks in different mobile network generations was outlined, including the post-5G technologies. This work provided also some security solutions to protect the 5G system against these threats, involving securing the data information exchange over the radio link and make the access more difficult for malicious parties.

Unfortunately, these few research works are still not enough to address the damaging 5G signaling threats, involving the DoS signaling attack tackled in this work. Hence, this paper extends our defense mechanism proposed in [10], as a preventive solution to defend against DoS signaling attack in 3G network, to meet also the problem of signaling threat in 5G system. Based on mMTC traffic model, the proposed mitigation mechanism based randomization technique has shown also promising results in decreasing the signaling load generated by the 5G infrastructure under signaling DoS attack while preventing the unnecessary use of the network resources.

The rest of the paper begins with a background section giving an overview of the new 5G RRC state model, and highlighting some security flaws of this novel RRC three-

state model. The section three analyses the 5G DoS signaling attack detection mechanism based randomization technique. This section presents first an overview on related works, then, it outlines the traffic model used for the performance evaluation of the detection framework, which is introduced at a later stage. Still in the same section, the simulation results are carried out to evaluate the effectiveness of the randomization based detection solution in defending against DoS signaling attack in 5G mobile network. Finally, the section four concludes the paper.

II. Background

In cellular systems, wireless communications between the devices and the network are carried out using the RRC protocol that is responsible for allocating and releasing the necessary radio resources. The signaling load produced by these resource allocation and release procedures will increase tremendously, specifically with the great variety of applications based on burst traffic (e.g., mMTC use case), which could disturb the proper functioning of the mobile networks infrastructures. As depicted in Fig. 1, in 5G system, a new RRC state, named $RRC_{INACTIVE}$, is introduced to meet the challenge of signaling overhead, battery life and latency. This novel $RRC_{INACTIVE}$ state is designed to reduce the latency by minimizing the signaling exchange triggered by the transition to RRC connected state $RRC_{CONNECTED}$ among the 5G infrastructure, which would be relevant for many smartphone applications that transmit small data on a frequent basis. This new state will also allow devices to conserve their batteries life by reducing the signaling load generated by the idle-connected states transitions. Indeed, in the $RRC_{INACTIVE}$ state, the device stores the RRC context (Access Stratum (AS) context) and maintains the core network (CN) connection established, and any detected traffic activity will trigger the transition to $RRC_{CONNECTED}$ state through a resume procedure using only three signaling messages instead of seven messages used in the switching process from the idle state (RRC_{IDLE}) to the connected state in 4G system [16]. The transitions between $RRC_{CONNECTED}$ and $RRC_{INACTIVE}$ states occur transparently to the CN. indeed, the CN network may carry any downlink traffic to the RAN entity so that the state transition from $RRC_{INACTIVE}$ to $RRC_{CONNECTED}$ does not involve any CN signaling exchange. As illustrated in Fig. 1, the new 5G RRC state model involves three states, namely, RRC_{IDLE}, $RRC_{CONNECTED}$ and $RRC_{INACTIVE}$. In this RRC three-state model, the transition from RRC_{IDLE} to $RRC_{CONNECTED}$ will primarily occur during the first UE attaches to the network or as a fallback to a new RRC connection. Hence, this transition will hardly arise when compared to the transition from $RRC_{INACTIVE}$ to $RRC_{CONNECTED}$, and with the shorter inactivity timeouts managing this later transition [17], the signaling load remains important even if the number of exchanged signaling messages related to the 5G RRC three-state transitions is reduced by introducing the $RRC_{INACTIVE}$ state, specifically when the 5G NG-RAN network is under a DoS signaling attack. indeed, a malicious exploiting of this inactivity timeout will give arise to two DoS attack scenarios. The first scenario is similar to the signaling attack tackled in [20], which aims at affecting and compromising an important number of MTC devices, and forcing them to send periodic burst packets after the expiration

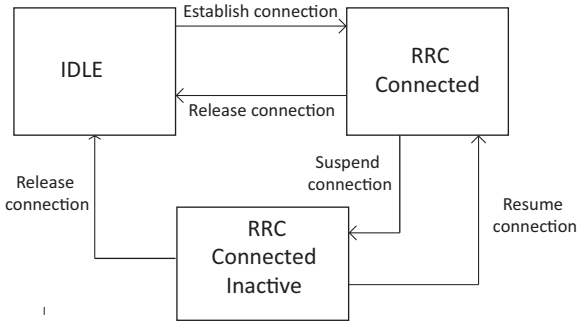


Fig. 1. 5G RRC State Machine Model.

of the inactivity timer to trigger frequent resource allocation and release procedures, thus causing a peak of signaling load that can not be properly sustained by the mobile infrastructure. Adversely, the second attack scenario aims to consume abusively the NG-RAN radio resources by maintaining a set of compromised devices in the $RRC_{CONNECTED}$ state for a considerable period of time leading to a network resource starvation. There are other security risks threatening the NG-RAN infrastructure, involving, the integration with the existing vulnerable systems, namely, Internet and 4G network, the immaturity of the 5G production process and maintenance procedures, and the overgrowth of the 5G components. These security flaws could amplify the risk of breaking down the confidentiality, integrity, and availability of network elements, and giving rise to more attack vectors against 5G system. Therefore, developing a robust a defense system that can protect the 5G system against such security threats, will be a serious challenge for mobile service providers.

III. 5G DoS Signaling Detection Mechanism based Randomization Technique

In this paper, we will evaluate the proposed detection mechanism based randomization approach regarding the DoS signaling attack exploiting the new 5G RRC three-state machine by analyzing the decreased signaling overhead ratio DSO_R and the network resource occupation time ratio ROT_R related to NG-RAN RRC handling process regarding different statistical distributions, namely, Gaussian, Log-normal and Exponential distributions. To carry out the performance evaluation of the proposed detection framework within 5G system, we will use the mMTC massive sensors traffic pattern [18] as 5G networks are expected to handle an significant amount of mMTC communications.

A. Related Works

To consolidate the security perimeter against signaling attacks in mobile networks, several protection mechanisms have been proposed in the literature review, specifically, for 3G and 4G networks. Among these defense solutions, a randomization method applied to some configuration parameters, like the channel inactivity timeout, has been proposed in [8], [9], [10], to increase the difficulty of hacking the value of such extremely vital network settings. According to [8], the randomization technique attributes the same random inactivity timeout to

all UEs handled by the same 3G Radio Network Controller (RNC) regardless of the traffic volume handled by these UEs. The randomization approach proposed in [8] presented some drawbacks related to a rise in resource consumption due to system configuration that becomes dynamic and no longer optimal, leading to an unbalanced resource consumption among different traffic patterns. Hence, [10] has proposed an enhanced randomization based detection framework to cope with the DoS signaling attacks in 3G system while optimizing the resulting resource consumption as well. Indeed, this improved randomization technique deployed an additional concept of classifying the devices according to the traffic volume periodically received by the 3G control plane over the corresponding measurement reports. In 5G context, the randomization approach has been used to defend against paging message hijacking attack [13]. Indeed, this solution aims at randomizing the paging occasion, which consists on changing the paging occasion after every paging cycle regardless of whether the 5G device received any paging message in that paging cycle. Such an approach, however, depletes rapidly the available P-TMSI values, and requires that the device and the base-station should be accurately synchronized.

B. Traffic Modeling: mMTC use Case

mMTC communications connect a plenty of devices constrained by cost and energy considerations. mMTC can be used for monitoring and area-covering measurements through sensor and actuator deployments. This 5G traffic use case is usually modeled using the 3GPP bursty traffic FTP model 3 [18], which is based on Bursty traffic with a fixed-size packet following a Poisson arrival process with rate λ, packet inter-arrival time $f_{D,mMTC}(t)$ and packet size $f_{Y,mMTC}(t)$. According to [18], the number of mMTC devices is about 25000 per cell, in this paper, we will simulate the traffic pattern related to N_{mMTC} connected devices. Using the traffic model parameters described in Table I, we will first simulate the mMTC signaling load generated by the new 5G RRC state handling under a different DoS signaling attack scenarios in accordance with various $T_{5G_{inac}}$, namely, 1 s, 2 s and 3 s. Then, we will evaluate the DSO_R and ROT_R metrics to demonstrates the effectiveness of the proposed defense solution in mitigating the DoS signaling attack in 5G system.

TABLE I. MMTC SIMULATION PARAMETERS

N_{mMTC}	T_s	T_{inac}	$f_{Y,mMTC}(t)$	$f_{D,mMTC}(t)$
1000	7200 s	1 s, 2 s, and 3 s	125 B	1 s

C. Detection Framework

For the mMTC traffic model, we have a well known behaviour of devices, which transmit the same amount of data $f_{Y,mMTC}$ during a defined transmission time period $f_{D,mMTC}$, so the data traffic classification is meaningless in this case. To this end, we will use the randomisation techniques as follows:

For the Gaussian distributions, μ is set to $T_{5G_{inac}}$, and $\sigma = T_R$.

For the exponential case, we use a modified exponential distribution (weighted by a factor w), the λ are computed as:

$$\begin{cases} \frac{1}{\lambda} = T_{5G_inac} \times w; \quad w = \sqrt{\frac{{T_{5G_inac}}^2}{T_R}} \end{cases} \tag{1}$$

For the log-normal distribution, the μ and the σ are computed as follows:

$$\begin{cases} \mu = \log(\frac{{T_{5G_inac}}^2}{\sqrt{T_R + {T_{5G_inac}}^2}}) \\ \sigma = \sqrt{log(\frac{T_R}{{T_{5G_inac}}^2} + 1)} \end{cases} \tag{2}$$

Where:

$$\{ T_R = a * T_{5G_inac} \tag{3}$$

The weighted parameter a is set so that the available inactivity timers remain in the interval [1s 10s] defined for 5G standard.

D. Analysis and Results

To evaluate the performance of the proposed detection mechanism, we will analyze two metrics, namely, the DSO_R related to the promotion state transition to $RRC_{CONNECTED}$, and the ROT_R which refers to the ration of time period that device remains inactive in $RRC_{CONNECTED}$ state in normal case ($T_{5G_{inac}}$ is static) regarding the resource occupation time related to randomized $T_{5G_{inac}}$.

$$DSO_R = \frac{SL(N) - SL(R)}{SL(N)} \tag{4}$$

$$ROT_R = \frac{T_{\mathrm{RO}}(N) - T_{\mathrm{RO}}(R)}{T_{\mathrm{RO}}(N)} \tag{5}$$

Where:

$$\begin{cases} SL\text{: Signaling Load (in number of signaling messages)} \\ T_{\mathrm{RO}}\text{: Resource Consumption Time} \\ R\text{: Randomization} \\ N\text{: Normal case} \end{cases}$$

By periodically launching a DoS signaling attack using different numbers of compromised mMTC devices (10%, 25% and 50% of the total number of simulated devices N_{mMTC}) every $T_{5G_{inac}}$ (attack period), we will first evaluate the generated signaling load when no defense mechanism is implemented for different inactivity timeouts, namely, 1s, 2s and 3s. From the simulation results depicted in Fig.2, Fig. 3 and Fig. 4, we can infer that the mMTC traffic pattern gives rise to a larger signaling load for the smaller inactivity timers even in case when no DoS signaling attack is initiated. The high amount of signaling traffic for the small value of inactivity timer ($T_{5G_{inac}}$ =1s) can be justified by the fact that the mMTC traffic pattern

is a Poisson distribution with a mean inter arrival rate λ_{mMTC} about one packet per second, thus, a higher $T_{5G_{inac}}$ (superior to 1s) means less state transitions between $RRC_{INACTIVE}$ and $RRC_{CONNECTED}$ states and then less signaling load.

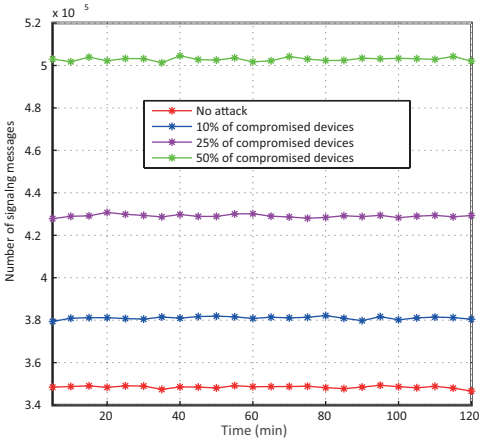


Fig. 2. New 5G RRC Model Signaling Overhead under a DoS Signaling Attack for $T_{5G_{inac}} = 1$s.

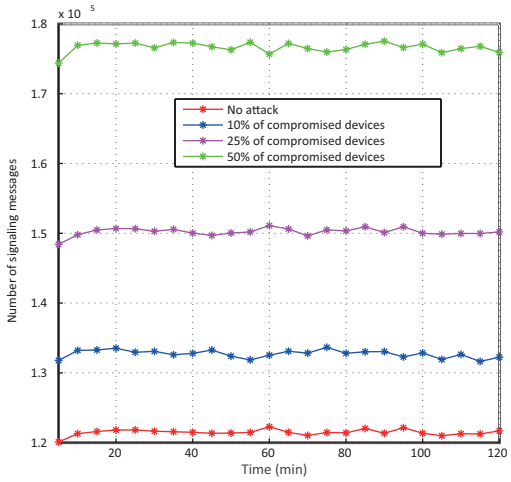


Fig. 3. New 5G RRC Model Signaling Overhead under a DoS Signaling Attack for $T_{5G_{inac}} = 2$s.

Regarding the two simulated metrics DSO_R and ROT_R, the performance evaluation of the randomisation based detection mechanism has shown promising results in mitigating the impact of DoS signaling attack against the novel 5G RRC three-state model. As illustrated in Fig. 5 and Fig. 6, the three simulated distributions, namely Gaussian, Log-normal and exponential functions reduce considerably the signaling overhead and the unnecessary resource consumption, which reach 70% and 65%, respectively for the exponential distribution with $T_{5G_{inac}} = 1$s and 50% of compromised mMTC devices. We have choose to evaluate our detection mechanism regarding the $T_{5G_{inac}} = 1$s, due to the large volume of signaling load generated by using this smaller inactivity timer, which constitute the most devastating attack scenario, specifically by compromising 50% of total mMTC devices.

As outlined in Table II, the randomization technique has shown better results in 5G context when compared to 3G

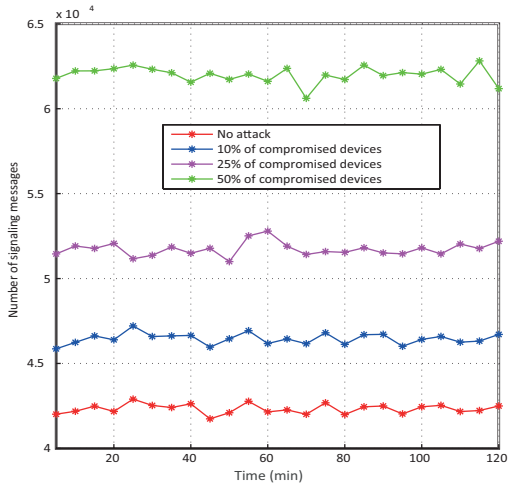


Fig. 4. New 5G RRC Model Signaling Overhead under a DoS Signaling Attack for $T_{5G_{inac}} = 3$s.

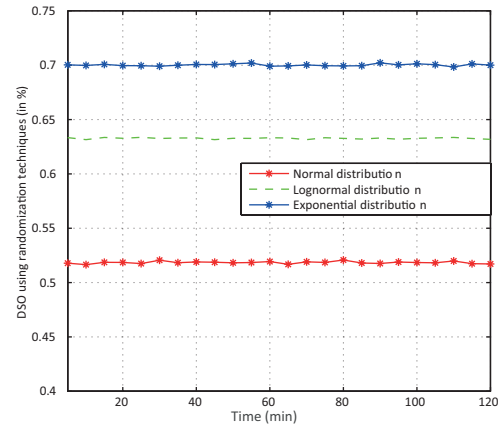


Fig. 5. Decreased Signaling Overhead using Randomization Techniques for 50% of Compromised mMTC Devices: $T_{5G_{inac}} = 1$s.

context, specifically for the Exponential and Log-normal distributions. Hence, the randomization approach remains very promising solution to be considered in mitigating the signaling threats in new mobile network generations. First, this technique offers a preventive framework that can avoid the occurrence of such attacks or at least mitigate their impact. Secondly and from a hardware perspective, the proposed randomized approach needs simply some low-complexity software updates in only some network entities.

IV. Conclusion

In this paper, we have extended our detection mechanism based randomization technique to defend against DoS signaling attack emerged in the new 5G RRC three-state model. The proposed solution has shown promising results in mitigating the impact of these signaling threats in 3G system, and we have demonstrated through simulation based on mMTC traffic pattern, the effectiveness of our detection framework regarding the 5G system as well. Indeed, for an inactivity timeout equal to 1 s and 50% of compromised mMTC device, the three simulated randomisation methods decrease significantly

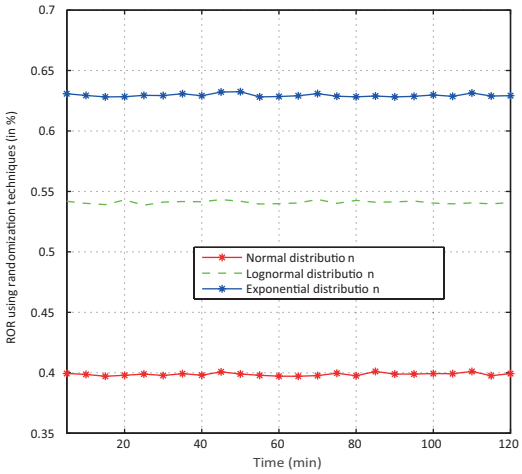


Fig. 6. Resource Consumption Ratio using Randomization Techniques for 50% of Compromised mMTC Devices: $T_{5G_{inac}} = 1$s.

TABLE II. RANDOMIZATION BASED DETECTION FRAMEWORK: 5G VS 3G PERFORMANCE COMPARISON

Randomization approach	3G system [10]		5G system	
	DSO_R (%)	ROT_R (%)	DSO_R (%)	ROT_R (%)
Gaussian distribution	46.53	55.99	52	40
log-normal distribution	42.27	10.98	64	54
Exponential distribution	31.91	13.12	70	64

the signaling load while avoiding the unnecessary network resource use. For the exponential distribution, the decreased signaling load is up to 70%, and the resource consumption ratio is around 65%, which constitutes an significant enhancement of network performances concerning the signaling overhead and the resource starvation raised from the new 5G designs, specifically when the network is under a DoS signaling attack. Our future work revolves around deeper analysis of new emerging signaling threats in the next generation (NG) of mobile systems, and new proposals to build a robust detection mechanisms to defend against the signaling attacks.

REFERENCES

[1] ITU-R, "IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond," Recommendation M.2083-0, September 2015.

[2] 3GPP, "3GPP TSG RAN WG1 Meeting 87," November 2016.

[3] Da Silva, I. L., Mildh, G., Säily, M., & Hailu, S. (2016, May). A novel state model for 5G radio access networks. In 2016 IEEE International Conference on Communications Workshops (ICC) (pp. 632-637). IEEE.

[4] Lee, P. P., Bu, T., & Woo, T. (2007, May). On the detection of signaling DoS attacks on 3G wireless networks. In IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications (pp. 1289-1297). IEEE.

[5] Kambourakis, G., Kolias, C., Gritzalis, S., & Hyuk-Park, J. (2009, June). Signaling-oriented DoS attacks in UMTS networks. In International Conference on Information Security and Assurance (pp. 280-289). Springer, Berlin, Heidelberg.

[6] Pavloski, M. (2018, February). Signalling attacks in mobile telephony. In International ISCIS Security Workshop (pp. 130-141). Springer, Cham.

[7] Abdelrahman, O. H., & Gelenbe, E. (2014, June). Signalling storms in 3G mobile networks. In 2014 IEEE international conference on communications (ICC) (pp. 1017-1022). IEEE.

[8] Wu, Z., Zhou, X., & Yang, F. (2010, September). Defending against DoS attacks on 3G cellular networks via randomization method. In 2010 International Conference on Educational and Information Technology (Vol. 1, pp. V1-504). IEEE.

[9] Chandra, M., Kumar, N., Gupta, R., Kumar, S., Chaurasia, V. K., & Srivastav, V. (2011, April). Protection from paging and signaling attack in 3G CDMA networks. In 2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (pp. 406-410). IEEE.

[10] Ettiane, R., Chaoub, A., & Elkouch, R. (2016, October). Enhanced traffic classification design through a randomized approach for more secure 3G mobile networks. In 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 116-121). IEEE.

[11] Jover, R. P., & Marojevic, V. (2019). Security and protocol exploit analysis of the 5G specifications. IEEE Access, (Vol.7, pp. 24956-24963).

[12] Khan, R., Kumar, P., Jayakody, D. N. K., & Liyanage, M. (2019). A Survey on Security and Privacy of 5G Technologies: Potential Solutions, Recent Advancements, and Future Directions. IEEE Communications Surveys & Tutorials, 22,(Vol.1, pp. 196-248).

[13] Singla, A., Hussain, S. R., Chowdhury, O., Bertino, E., & Li, N. (2020). Protecting the 4G and 5G cellular paging protocols against security and privacy attacks. Proceedings on Privacy Enhancing Technologies, 2020, (Vol.1, pp.126-142).

[14] Hussain, S. R., Echeverria, M., Chowdhury, O., Li, N., & Bertino, E. (2019, February). Privacy Attacks to the 4G and 5G Cellular Paging Protocols Using Side Channel Information. In NDSS (Vol. 19, pp. 24-27).

[15] Ahmad, I., Shahabuddin, S., Kumar, T., Okwuibe, J., Gurtov, A., & Ylianttila, M. (2019). Security for 5G and beyond. IEEE Communications Surveys & Tutorials, 21, (Vol.4, pp. 3682-3722)

[16] SILVA, D.,MILDH, G., PAUL, S-B., MAGNUS, S.,& ALEXANDER, V.(19 June 2019). Meeting 5G latency requirements WITH INACTIVE STATE. ERICSSON TECHNOLOGY REVIEW.

[17] 4G-5G Interworking RAN-level and CN-level Interworking. White Paper, June 2017

[18] 5G PPP use cases and performance evaluation models. White Paper, v1.0, 2016. [retrieved: 2017-07-28].[Online]. Available: https://5g-ppp.eu/wp- content/uploads/2014/02/5G-PPP- use-cases-and-performance- evaluation-modeling v1.0.pdf

[19] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Licensed-Assisted Access to Unlicensed Spectrum;(Release 13)

[20] Ettiane, R., Chaoub, A., & Elkouch, R. (2018, May). Robust detection of signaling DDoS threats for more secure machine type communications in next generation mobile networks. In 2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON) (pp. 62-67). IEEE.

Regulation Proposal for the Implementation of 5G Technology in Peru

Luis Nuñez-Tapia
Facultad de Ciencias e Ingenieria
Universidad de Ciencias y Humanidades
Lima, Peru

***Abstract*—Telecommunications play a very important role in people's life, for years there has been an evolution of this technology in the mobile communications' industry reaching up to the 5G technology, which is more advanced than 4G, making it thus more comfortable for the user. In Peru, 5G technology has not been implemented though because there is fear from great part of population regarding its the antennas. Another fear is nowadyas the spread of COVID–19, this is because there is a lot of false information that has poor scientific support, even that information has been denied by the Ministry of Transport and Communications (MTC) from Peru, but still people hold on to these fears. Due to the aforementioned reasons, the present investigation aims to carry out an assessment of the benefits that the 5G technology would bring to the country and also proposes a regulatory frame for the radioelectric spectrum that will occupy this technology in Peru. By evaluating a regulation proposal of 5G technology in Peru, it is shown that the implementation of this technology will bring benefits in the social and economic sectors of the country.**

***Keywords*—5G; regulation; antennas; radio spectrum**

I. Introduction

The 5G technology has high expectations regarding its launch. If we observe the evolution that the mobile technology has had since its inception, it has shown great advancement. So it was expected that the 4G technology in some moment was going to be surpassed. And so, the technology 5G appeared in which people can navigate at higher speeds, that is, allowing individuals to have an internet connection faster than 4G, with the promise of 10 Gbps connectivity and latency less than five milliseconds ($<$ 5ms). Hence, it is no surprise that the current increase in the demand for mobile connectivity is going to accelerate dramatically [1].

In Peru, the 5G technology has not been incorporated due to missing antennas and base stations that can support this technology. According to the Ministry of Transport and Communications (MTC), there is no 5G antenna in the country at the moment nor it is not known if the implementation of them is being carried out during the stage of social confinement due to the COVID–19 pandemic in the country [2]. On the other hand, it has been observed that many citizens fear antennas and even relate it to the spread of COVID–19, this is because they get carried away by false information. Faced with abundant information from unreliable sources, the MTC through several statements and campaigns has indicated to the population that there is no relationship between 5G technology and the spread of COVID–19 [3]. According to the director of Audits and Sanctions in Communications, Patricia Daz, she has indicated that there is no scientific research to support in no way that the antennas have any relation also to cancer and/or any other disease [2].

In the South American nation of Peru there are people who distrust the antennas and even refuse to have them installed within their territory. For instance in the tiny Peruvian town of Chopcca (Huancavelica), villagers burned antennas and kidnapped the personnel in charge of the installation. This is really contradictory, as many people complain about the little coverage that there exist specially in the rural communities. Furthermore due to the current pandemic many students are not able to receive their classes remotely. They complain that the current internet connection they have is very slow and sometimes the image they see in video is pixelated. Hence, a paradox that stands out is that users want greater connectivity but they don't want any more antennas.

Some studies about the 5G technology, such as the work developed by [4], cover the benefits that 5G technology will generate, for example, in Ecuador. It has been indicated that this technology will take time to develop within this country since for the government it represents additional costs and even the operators do not have the necessary resources to do the tests. It should be noted that previously when a new technology such as the 4G has been implemented in Ecuador, there has been a growth in the economic, technological and social sectors. Another work developed in Indonesia [5] mentioned the importance of the features about the 5G technology. To conduct an evaluation and see if such technology can be integrated into this Asiatic country, the authors considered that this technology should be assigned to frequency spectrum that is already being occupied by fixed satellites within this country. The authors in the aforementioned study recommended performing a cost benefit analysis to know if it is convenient to use 5G technology in the selected spectrum of frequency. Finally there is a work developed in Colombia by [6], where a technology comparison between 4G and 5G is made. For this the authors made a description about the evolution that each technology has had in mobile networks and also the increase in their web browsing speeds. The authors of this Colombian study mentioned that the 5G technology will be beneficial in the social and economic realm.

Seeing the problems in Peru regarding the little knowledge that people have about the 5G technology, this article will conduct an assessment based on the benefits this technology will bring. In Section II, characteristics of the 5G technology will be presented, such as the concept of this technology and its radio spectrum radio for 5G. In Section III, we will show a

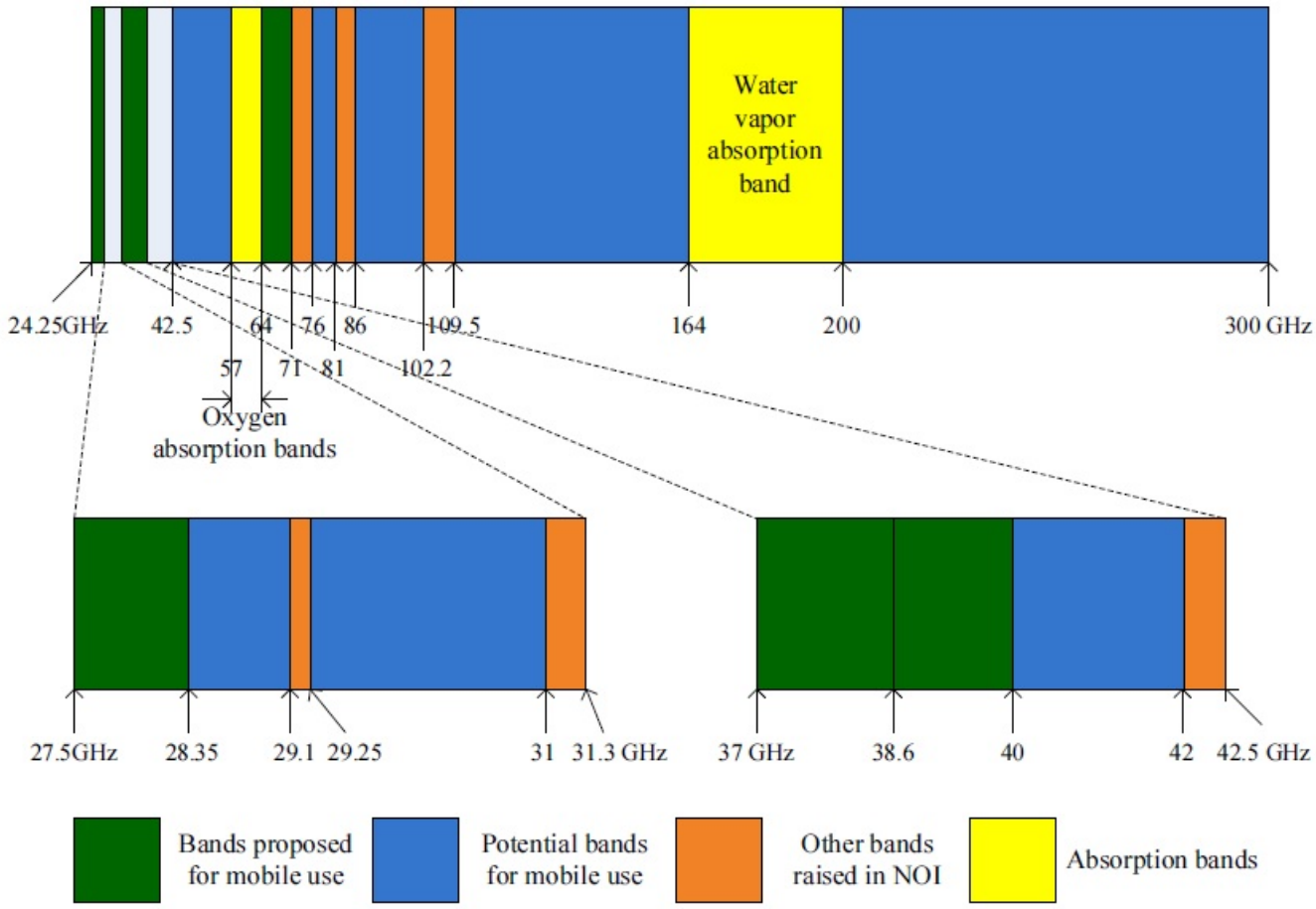


Fig. 1. Spectrum Usage in mmWave Bands [7].

brief description of the mobile telephony industry in Peru and also discuss the improvement that 5G will bring by considering the social and economic perspectives. Finally, in Section IV the conclusions are indicated.

II. Characteristics of 5G

A. Concept of 5G Technology

The 5G technology or the fifth generation one, is an evolution of mobile communications systems that throughout the years has been cherished, where each generation is distinguished in improved data transmission speeds. This technology features a data transmission speed of several gigabits per second, latencies of approximately 1ms and reduced energy consumption for wireless broadband [8]. According to these characteristics, the increase in internet connection speed is remarkable. On the other hand, this technology also provides solutions for automation, power, agriculture, among other applications [4]. This is important since it allows that communication between man and machine is fluent, and also spares the need that individuals physically control machines.

B. Radio Spectrum for 5G Technology

The 5G technology offers faster connection speeds to the internet, so it requires a greater radio spectrum and an efficient one. Since the radio spectrum is a finite resource, the International Telecommunications Union (ITU) that is the global agency responsible for the spectrum management of radio frequencies and the resources of the satellites in orbit [9]. The ITU has to regulate a radio spectrum that does not generate conflict with those technologies that work already in a certain frequency spectrum and can be maintained in such a way also that 5G can meet the expectations required.

For the 5G technology, decisions about the management of the spectrum will play a critical role in meeting of the expectations established for the 5G networks [9]. This technology will have a wide bandwidth so it can offer a capacity and speed of data transmission optimally for the comfort of the users. One of the novelties of 5G is that a frequency band that was previously not considered for mobile communications such as the 24 GHz band and it is seen in Fig. 1 can be used. So 5G implies a higher spectrum and which is widely available in the millimeter range. This can accommodate as well the Massive-MIMO implementation that involves multiple small antennas and processing in devices [10]. MIMO or multiple input and multiple output, is used to improve wireless communication and will result important for the 5G technology.

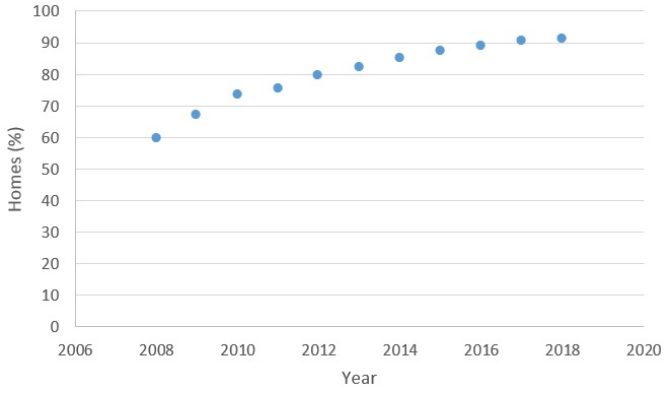


Fig. 2. Households with at least One Member of the Family having a Mobile Phone in the 2008-2018 period.

III. 5G in Peru

A. Mobile Communications in Peru

In Peru, the radioelectric spectrum being a finite resource, is administered by the Ministry of Transport and Communications (MTC). It is known that in various parts of the Peruvian territory there is no coverage, thus this limits the use of the maximum data transmission speeds that it already has for technologies such as 3G or 4G. The number of users has noticeably incremented in the last few years [11], however the technological capacity to satisfy all users has not increased.

At present there is a growth in mobile technology. Fig. 2 shows the number of households between 2008 to 2018 who have a mobile phone in Peru according to the National Institute of Statistics and Informatics from Peru. There is no doubt that a clear increment can be seen. Thus, this will lead to a growing demand of the radioelectric spectrum. To ensure that the development of 5G technology is optimal, a good management is needed, since the development of 5G involves large challenges at the technological and structural level. The architectures of mobile networks will have a notable change in their components, in the way of managing resources, and in the provision of services [12]. While it is true, some operators have conducted tests in Peru with 5G technology obtaining good results, only the tests were carried out and still this technology has not been implemented anywhere in the country.

The objective of implementing 5G technology is that all people can use it on their mobile devices to stay connected and being informed of what may happen in the world in a different way, without the famous bottleneck that occurs when a large number of users use the network at the same time. Therefore, the management of the radio spectrum for 5G technology is a crucial matter. On the other hand, the Supervisory Agency Private Organism in Telecommunications (OSIPTEL) from Peru, mentions that the arrival of these technologies require highly trained professionals to develop it [13].

B. Social Perspective

Since the 5G technology has not been implemented in Peru, the social perspective will not be so exact, so the analysis to know how the Peruvian population would take 5G technology could be evaluated by observing the impact the 4G technology has generated in the society of the country, and to what extent this has improved the comfort of the users.

The 4G technology in Peru has been well received, because it was a different experience from its predecessor. Its greatest internet browsing speed is of 100Mbps, communication of people by video call, etc.. Thus, this has generated greater expectations about the mobile communications that 5G technology will bring about. Currently with the pandemic generated by COVID–19, many people telecommute, students have classes remotely, so it should come as no surprise that when this technology is implemented in the country, it will increase the consumption of this one, since most of the users will always seek comfort. Although there are people who oppose the implementation of this technology due to the abundant false information that exists, the truth is that there is not yet scientific support that confirms the damage that 5G may generate.

C. Economic Perspective

To know what benefits the 5G technology in the country will have from an economic perspective, this could be evaluated using the growth that mobile devices have had. While it's true, mobile devices have grown considerably, now with the arrival of the 5G technology there is no doubt that growth will be greater. Moreover, if proper policies are implemented, the use of 5G technology in Peru could give a kick to the e-government and e-commerce sectors of the country. It should be noted that for the implementation of this technology in the country, it would have to make some regulations with the radioelectric spectrum. Although in the country, the spectrum they propose is being used for mobile communications, so there won't be much inconvenient for the use of 5G technology. The implementation of the 5G technology in the country, initially will be carried out in places where there are a greater number of users, that is to say where the industries that move the economy in the country are centralized. It has to be indicated that 5G implementation for the whole country would be really slow due to the lack of technology at the moment. In Fig. 3 we can observe both perspectives considered for the regulation of the implementation of the 5G technology in Peru.

IV. Conclusions

From the present investigation, it is concluded that the implementation of 5G technology in the country will bring benefits in the social and economic sectors. Mainly because users will have the convenience that this technology will provide. Furthermore, the demand of mobile telephony in the country will have a growth each year. This is because users always seek to have a tool that makes things easier for them. As future work, an analysis of this technology considering mathematical models in order to know exactly how much revenue will generate in the economy of Peru will be considered.

Acknowledgment

The author is very thankful to Dr. Carlos Sotomayor-Beltran for his insightful comments and helpful suggestions in improving this article.

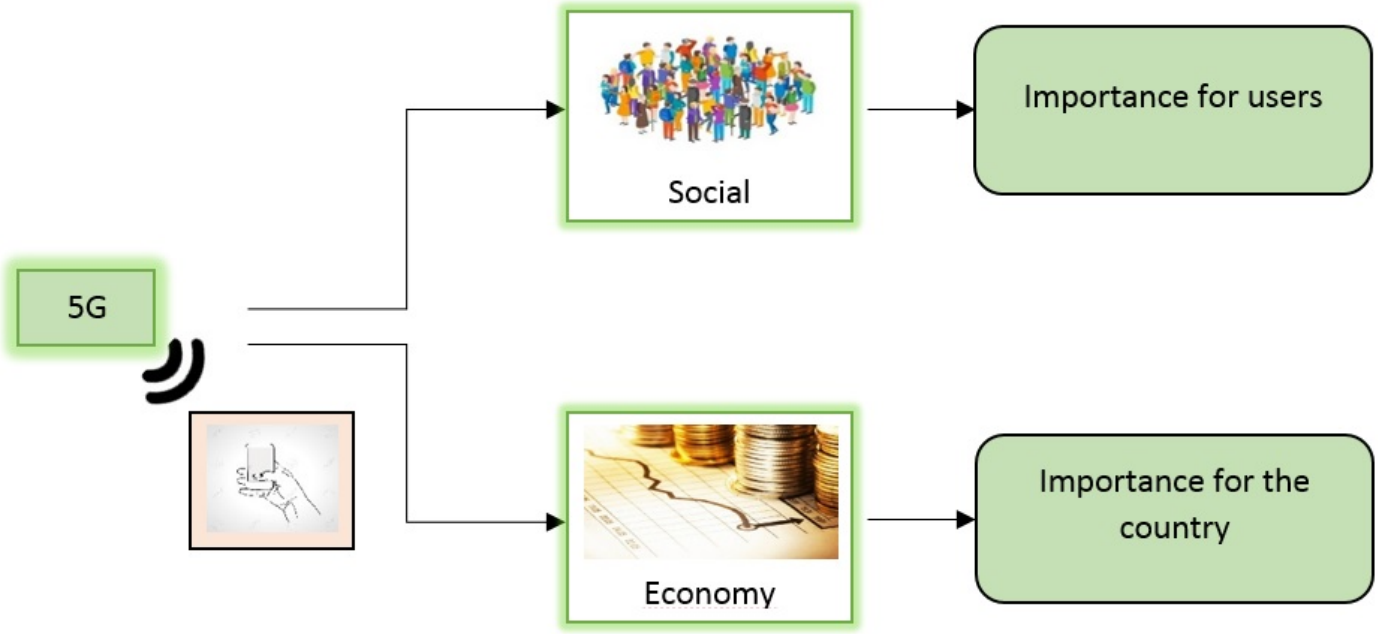


Fig. 3. Social y economic Perspective for the Implementation of the 5G Technology in Peru.

REFERENCES

[1] L. Elizabeth CondeZhingre, P. A. QuezadaSarmiento, and M. Labanda, "The new generation of mobile networks: 5g technology and its application in the eeducation context," in *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, 2018, pp. 1-4.

[2] RPP. (2014) "basta ya de mitos": Mtc confirma que no hay antenas 5g en el país". [Online]. Available: https://rpp.pe/tecnologia/mas-tecnologia/antenas-5g-en-peru-mtc-confirma-que-no-hay-infraestructura-5g-en-el-pais-y-no-hay-relacion-entre-coronavirus-y-la-red-covid-19-noticia-1272765?ref=rpp

[3] Ministerio de Transportes y Comunicaciones (MTC). (2020) Comunicado sobre información falsa. [Online]. Available: https://www.gob.pe/institucion/mtc/noticias/208626-comunicado-sobre-informacion-falsa

[4] A. Ulloa, "Estudio de la tecnología 5g y el impacto que tendrá en el país," Universidad de Guayaquil, 2018.

[5] L. Sastrawidjaja and M. Suryanegara, "Regulation challenges of 5g spectrum deployment at 3.5 ghz: The framework for indonesia," in *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, 2018, pp. 213–217.

[6] F. Guevara, "Comparativo entre la tecnología de redes 4g y 5g y los beneficios de su implementación en colombia," Universidad Santiago de Cali, 2018.

[7] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, C. I, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1909-1935, 2017.

[8] D. N. ArizacaCusicuna, J. Luis ArizacaCusicuna, and M. Clemente-Arenas, "High gain 4x4 rectangular patch antenna array at 28ghz for future 5g applications," in *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 2018, pp. 1-4.

[9] International Telecommunication Union (ITU). (2019) Itu-r: Managing the radio-frequency spectrum for the world. [Online]. Available: https://www.itu.int/en/mediacentre/backgrounders/Pages/itu-r-managing-the-radio-frequency-spectrum-for-the-world.aspx

[10] F. Saccardi, L. Scialacqua, A. Scannavini, L. J. Foged, L. Duchesne, N. Gross, F. Herbiniere, P. O. Iversen, and R. Braun, "Accurate and efficient radiation test solutions for 5g and millimeter wave devices," in *2018 IEEE MTTS Latin America Microwave Conference (LAMC 2018)*, 2018, pp. 1-4.

[11] C. Sotomayor-Beltran and L. Andrade-Arenas, "A spatial assessment on internet access in peru between 2007 and 2016 and its implications in education and innovation," in *2019 IEEE 1st Sustainable Cities Latin America Conference (SCLA)*, 2019, pp. 1–4.

[12] S. De Los Ríos and D. Quiñónez, "5g: An holistic, systemic and prospective vision," in *2017 IEEE Colombian Conference on Communications and Computing (COLCOM)*, 2017, pp. 1-6.

[13] Organismo Supervisor de Inversión Privada en Telecomunicaciones (OSIPTEL). (2019) Nuevos retos en el sector telecomunicaciones exigen capital humano altamente capacitado. [Online]. Available: https://www.osiptel.gob.pe/noticia/np-nuevos-retos-sector-telecom-capital-humano-capacitado

A Meta-analysis of Educational Data Mining for Predicting Students Performance in Programming

Devraj Moonsamy[1], Nalindren Naicker[2], Timothy T. Adeliyi[3], Ropo E. Ogunsakin[4]
Department of Information Systems, Durban University of Technology, Durban, South Africa[1, 2, 4]
Department of Information Technology, Durban University of Technology, Durban, South Africa[3]

***Abstract*—An essential skill amid the 4th industrial revolution is the ability to write good computer programs. Therefore, higher education institutions are offering computer programming as a module not only in computer related programmes but other programmes as well. However, the number of students that underperform in programming is significantly higher than the non-programming modules. It is, therefore, crucial to be able to accurately predict the performance of students pursuing programming since this will help in identifying students that may underperform and the necessary support interventions can be timeously put in place to assist these students. The objective of this study is therefore to obtain the most effective Educational Data Mining approaches used to identify those students that may underperform in computer programming. The PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analysis) approach was used in conducting the meta-analysis. The databases searched were, namely, ACM, Google Scholar, IEEE, Pro-Quest, Science Direct and Scopus. A total of 11 scientific research publications were included in the meta-analysis for this study from 220 articles identified through database searching. The residual amount of heterogeneity was high ($\tau 2 = 0.03$; heterogeneity $I^2 = 99.46\%$ with heterogeneity chi-square = 1210.91, a degree of freedom = 10 and P = <0.001). The estimated pooled performance of the algorithms was 24% (95% CI (13%, 35%). Meta-regression analysis indicated that none of the moderators included have influenced the heterogeneity of studies. The result of effect estimates against its standard error indicated publication bias with a P-value of 0.013. These meta-analysis findings indicated that the pooled estimate of algorithms is high.**

Keywords*—Data mining; educational data mining; machine learning; performance; programming

I. Introduction

An essential skill not only in IT programmes in higher education but other disciplines as well is the ability to write good computer programs. However, the failure rate of programming relative to other subjects that students pursue is significantly higher [1]. Furthermore, we are currently in the 4th industrial revolution and it is imperative that graduates acquire this important skill to add value to the organizations that will employ them in the future. It is therefore important to be able to predict the performance of students wanting to pursue programming to put in place the necessary interventions for students that are likely to underperform in programming. The prediction of students' performance in programming can therefore be facilitated through the process of Educational Data Mining.

In the not too distant past data analysis was performed using mathematical and statistical methods using tools like charts, regression methods etc. to assist in decision making. However, because the amount of information in the world is increasing very rapidly coupled with an increase in the number of databases, the production of useful information has become very challenging and primitive tools can no longer be used in the analysis of these huge data sets. The type of analysis that needs to be performed on the data to extract interesting, important and meaningful patterns of information thereby allowing its applicability in many areas of our lives is called Data Mining (DM) [2-4].

Data Mining (DM) is also known as Knowledge Discovery from Data (KDD) which converts enormous amounts of data into knowledge. In DM data is explored from different perspectives to derive useful information from the data [5]. Closely related to Data Mining is Educational Data Mining and as illustrated by Ventura et al. in [6] Educational Data Mining shares many attributes from other disciplines like education, computer science and statistics [5, 7-12].

Educational Data Mining (EDM) attempts to obtain knowledge from educational data by building models to facilitate the examination of educational data to discover important student related information [5]. Educational Data Mining is a relatively new discipline that employs various methods to extract meaning from huge amounts of data found in educational environments in order to better understand students' behaviour and results. The primary goal of EDM is to decipher how students learn and to identify those factors that will enhance students learning [13].

A desired outcome of EDM is to be able to predict the performance of students since this is closely related to the quality of education. The resulting prediction models created as an outcome of EDM can help educators identify problems faced by students that may be affecting their academic performance [1]. Numerous studies have been conducted in predicting the performance of students not necessarily in programming, including studies in [14-16].

This study is a meta-analysis of Educational Data Mining research with the aim of obtaining the most effective Educational Data Mining approaches used to predict the performance of students pursuing computer programming. Aligned to the aim the research question of this study is as follows: What are the most effective EDM methods used for prediction of student performance in computer programming? This paper consists of the following sections: Section II is a

discussion of related works about studies involving the prediction of students' performance in programming. Section III is a discussion of the methodology used. In Section IV the results and findings of the meta-analysis are presented. The limitations of the study are discussed in Section V and finally, the paper concludes in Section VI.

II. Related Works

Many studies have been conducted to predict students' performance in programming [1, 17, 18]. An analysis of the literature reveals that the studies conducted can be categorized into the following two broad categories namely, studies carried out to predict student performance in programming using their performance in a programming related module either at school or in a programming related entrance test; and studies conducted to predict student performance in programming using other features like background factors, grades obtained in mathematics or physical science or other factors not directly related to programming. In this section, the literature from these two perspectives are presented.

In research conducted by Sivasakthi in [19], five data mining algorithms were executed on a data set to predict students' performance in an introductory programming module. These algorithms were: Multilayer Perceptron, Naïve Bayes, SMO, J48 and REPTree. The study used student demographic related data, the grade obtained in programming at college (i.e. before university) and the grade obtained in an entrance test. It was found that MLP performed best with an accuracy of 93% and the Naïve Bayes algorithm had the lowest prediction accuracy of 84%. In the MLP method, the factor that lead to the highest prediction of students' performance was students' grade obtained in college and the entrance test. Because many students pursuing programming have not programmed previously be it at school or elsewhere, this model will not be able to predict the performance of students with no prior programming exposure.

Pathan et al. in [5] developed a DT model to classify C programming students into 3 groups good, average and poor. The attributes used in this study were related to student behaviour and past educational information as well as C programming questions. The DT model by Pathan et al. in [5] was able to classify 87% of students correctly.

In a study by Đambić et al. in [20] a machine learning model was developed to predict the likelyhood of students pursuing an entry level programming module of failing. The features that were used in the model are as indicated in Table I:

TABLE I. Features for Model

Feature	Description
X_1	Number of points from the first colloquium
X_2	Number of points from the first quiz
X_3	Number of points from the first homework
X_4	Whether is this a second-time student has enrolled in this course
X_5	Whether the student has attended the first colloquium

This study used the logistic regression model. The misclassification of the model was around 19% and the precision was around 67%. The use of this model simply meant that many students who would have passed on their own were identified and would be sent for additional support interventions.

Costa et al. in [21] attempted to determine the efficiency of four EDM techniques namely Decision Tree, Support Vector Machine (SVM), Neural Network and Naive Bayes. These techniques were implemented on two independent sets of data pertaining to entry level programming modules at a university in Brazil. The data sets were data from residential students and the other included data from distance education students. The study revealed that the SVM technique performed far better than the other EDM techniques by predicting with an accuracy of 92% for distance education students and with an accuracy of 83% for residential students.

Figueiredo et al. in [22] proposed a neural network predictive model for predicting student failure in programming using their performance in various programming related tasks during class. This model enabled teachers to filter out those students that are more likely to fail early enough to implement new teaching interventions so as to enhance the students programming skills. The neural network model had an accuracy of 94.12% and a precision of 95.45%.

Vihavainen et al. in [23] investigated how students programming behaviour (e.g. eagerness to work on programming exercises) influences their grade in the module. In this study, only data derived online taking screen shots of students programing exercises were used. Furthermore, students' background information was not used as features in this study. The study predicted with a 78% accuracy as to whether the student was a high-achiever, passed the module, or failed the module.

In the study by Bergin et al. in [24] six machine learning algorithms were considered in the prediction of student performance in programming. The study used several categories of predictors of performance in programming. The categories include background factors, factors related to comfort level at the commencement of the module (This category included programming related questions), motivation and the student use of learning techniques. Naïve Bayes out-performed the other machine learning algorithms by being able to predict with an accuracy of 78.3%.

Aguinaldo et al. in [25] developed a predictive model to determine student's success in an introductory programming module using six 21st century learning skills which are: Creative Skill, Reflective Skill, Problem- Solving Skills, Collaborative, Communication and Adaptability Skills. This predictive model used the PART classifier algorithm. It was found that communication was the strongest predictor of success in programming logic formulation. Unlike the study by Sivasakthi in [19] this predictive model was not based on performance in programming and can therefore be utilised to predict the performance of students who have no prior programming exposure.

In a study by Abdulsalam et al. in [2] three decision tree algorithms which are C4.5 (J48 in WEKA), CART and BF were used in predicting the performance of students in computer programming using the attributes of the grades obtained in Mathematics and Physics. The study revealed that J48 performed better than the CART and BF algorithm. J48 had a prediction accuracy of 70.37% while CART and BF Tree had prediction accuracies of 60.44% and 60.30% respectively. In a similar study conducted at a Nigerian university using a prediction model based on Artificial Neural Networks (ANN) it was also found that students possessing above average grades in Mathematics and Physics performed better in programming as compared to students who did not possess these attributes [26].

In the study by Mohamad et al. in [27] rough set was applied to a data set in order to identify those factors that influenced students' performance in programming based on data from earlier student results. The study revealed that students who have attempted a programming course before university and students who have obtained an average mark for mathematics, English and the Malay language at school were good indicators of performance in programming at university. In addition, in terms of personality factor, the investigative and social type student and the average cognitive student were identified as important attributes that effect the performance in computer programming.

Badr et al. in [28] developed a model to predict the performance of students wanting to pursue programming. This model used as attributes the marks that students obtained in mathematics and English. In this study, a classifier was built using an association rules algorithm. Unlike many other studies, this study resulted in the creation of a model that was able to predict a students' likelihood of success in programming before registering for the course. This meant that the performance of students pursuing programming increased since they could adjust their teaching strategies to accommodate those students that were predicted to more likely underperform in the programming course. The study conducted two experiments by executing the CBA rule-generation algorithm. The first used the marks obtained in English and mathematics modules, and this resulted in four rules with an accuracy of 62.75%. The second used marks obtained in only English, resulting in four rules with an accuracy of 67.33%.

Table II summarizes the various studies in the literature that used data mining or machine learning algorithms in the prediction of students' performance in programming. The table is classified according to the following headings namely: author, problem focus, scientific method, sample size, classification of the algorithms and accuracy.

TABLE II. SUMMARY OF STUDIES CONDUCTED TO PREDICT STUDENTS' PERFORMANCE IN PROGRAMMING

Author	**Problem Focus**	**Scientific Method**	**Sample size**	**Classification of Algorithms**	**Accuracy**
[25]	A creation of a Predictive Model using 21st Century Learning skills.	PART classifier -algorithm	180	Hybrid	
[22]	The development of a Neural Network (NN) model to predict student failure using the attributes of students' gathered during class activities and assessments.	Multiple Back-Propagation (MBP) algorithm,	85	Data Mining	**94.12%**
[21]	An investigation of the efficiencies of four educational data mining techniques used to envisage those students that may under perform in a programming module.	Neural Networks, Decision Trees (J48), Support Vector Machine (SVM) and Naive Bayes	161	Data Mining	**92%**
[19]	The application of data mining algorithms such as multilayer perceptron, Naive Bayes, SMO, J48, REPTree on student related data to determine those students that may require additional support.	Multilayer Perceptron, Naïve Bayes, SMO, J48 and REPTree Survey cum experimental methodology	300	Data Mining	**93%**
[28]	The development of a model to predict students' performance in a programming module based on their performance in other modules.	CBA algorithm	203	Data Mining	**62.75%.**
[20]	A model to identify students who might have problems passing an Introduction to programming course.	logistic regression, simple quadratic model.	181	Hybrid	**81%**
[24]	A study of six machine learning algorithms to determine student success in computer programming.	Naïve Bayes	26	Machine Learning	**78.3%**
[2]	A study to identify the optimal DT algorithms for determining students' success in programming.	C4.5 (known as J48 in WEKA), Classification and Regression Tree (CART), and Best-First Tree (BF Tree)	131	Data Mining	**70.37%**
[5]	The creation of a decision tree (DT) mining model for improving students programming ability in C.	DT	70	Data Mining	**87%**
[23]	An investigation into how students' behaviour during the programming process affects the course outcome.	Bayesian network classifier	200	Data Mining	**78%**
[27]	Rough set was applied to a programming data set in order to determine those factors that will influence students success in programming.	Rough set, clustering, and association rule	419	Data Mining	**90%**

III. Research Method

A. Literature Search Strategy

The study was carried out using the PRISMA (preferred reporting items for systematic reviews and meta-analysis) approach [29-31]. In conducting the meta-analysis, many databases were searched including ACM, Google Scholar, IEEE, Pro-Quest, Science Direct and Scopus. Only papers published in English between the period 2010 and 2020 were retrieved from the databases. The following combination of terms were used in searching the various databases: 'Programming' [All Fields] AND 'Machine learning' [All Fields] OR 'Programming' [All Fields] AND 'Data Mining' [All Fields] OR 'Programming' [All Fields] AND 'Intelligent Systems' [All Fields] OR 'Programming' [All Fields] AND 'Problem Solving' [All Fields] OR 'Programming' [All Fields] AND 'Higher Education' [All Fields]. The search terms were separated or combined using the Boolean operators "OR" or "AND". All papers identified by the search were imported into EndNote X9. A total of 220 articles were identified between the years 2010 and 2020 as indicated in Fig. 1 below. Furthermore, the reference lists of related articles were also manually checked for citations overlooked during the searching of the databases.

B. Inclusion Criteria

The inclusion criteria of the articles were that the studies were carried out at higher education institutions where the performance of students in programming using machine learning or data mining algorithms were studied.

C. Exclusion Criteria

Articles written in languages other than English, published before January 2010 were excluded. Systematic reviews, editorials, books, book chapters and thesis were excluded. Articles on the performance of students in programming at schools were also excluded. Studies related to performance prediction of students in subjects other than programming were also excluded.

D. Statistical Data Analysis

The appropriate principal studies data were obtained and then captured onto an Excel sheet, which facilitated it being exported to the statistical analysis software, STATA version 15. Furthermore, the study incorporated the use of forest plots to estimate pool effect size and the effect of each study with their confidence interval (CI) to provide a visual image of the data. In a meta-analysis, it is essential to assess heterogeneity between the pooled studies. Heterogeneity in a meta-analysis denotes the dissimilarity in the results of the various studies. The index of heterogeneity (I^2 statistic) was used to assess the heterogeneity amongst the included studies and we tested for its significance using Cochran's Q test [32-34]. The I^2 statistic is used to denote the percentage of disparity amongst the studies that is attributed to heterogeneity and not chance. The I^2 values of 25%, 50%, and 75% indicate low, medium, and high heterogeneity, respectively. The meta-analysis amongst the subgroups were conducted to assess the mean pooled performance estimates based on the different types of algorithms.

Publication bias refers to biasness that is found in published academic research. Publication bias happens when the results of an experiment or study effects the decision as to publish the study or distribute it. Thus, only publishing studies that show a noteworthy finding affects the outcome of the research findings. In addition, publication bias can also result in the formulation and testing of hypotheses that is based on incorrect perceptions from the scientific literature. Hence, in this study, small study effect and funnel plot test were evaluated to assess the risks of publication bias. Furthermore, publication bias was assessed by means of Egger's and Begg's test [35, 36].

As indicated in Fig. 1, this systematic review includes published papers between January 2010 and November 2020. These articles were then imported into EndNote version X9 and the duplicates removed, resulting in 196 articles remaining. A further 25 articles were removed after reading the abstracts. Following the review of the 171 articles, 139 articles were deleted due to various reasons and a further 21 excluded due to the specified inclusion and exclusion criteria. The smallest sample size was 26 participants in a study conducted with a machine learning algorithm, while the largest sample size was data mining algorithm approach. A total of 1956 participants were included in this meta-analysis. Most of the studies were carried out with the data mining algorithm approach, 8 (73%), hybrid algorithm, 2 (18%), and the remaining were performed with a machine learning approach, 1 (9%). When we look at the subgroup where the prediction was made, we found that three of the included studies was used to make a prediction and three on student-related prediction. Fig. 1 below illustrates the PRISMA approach used in conducting the database searches.

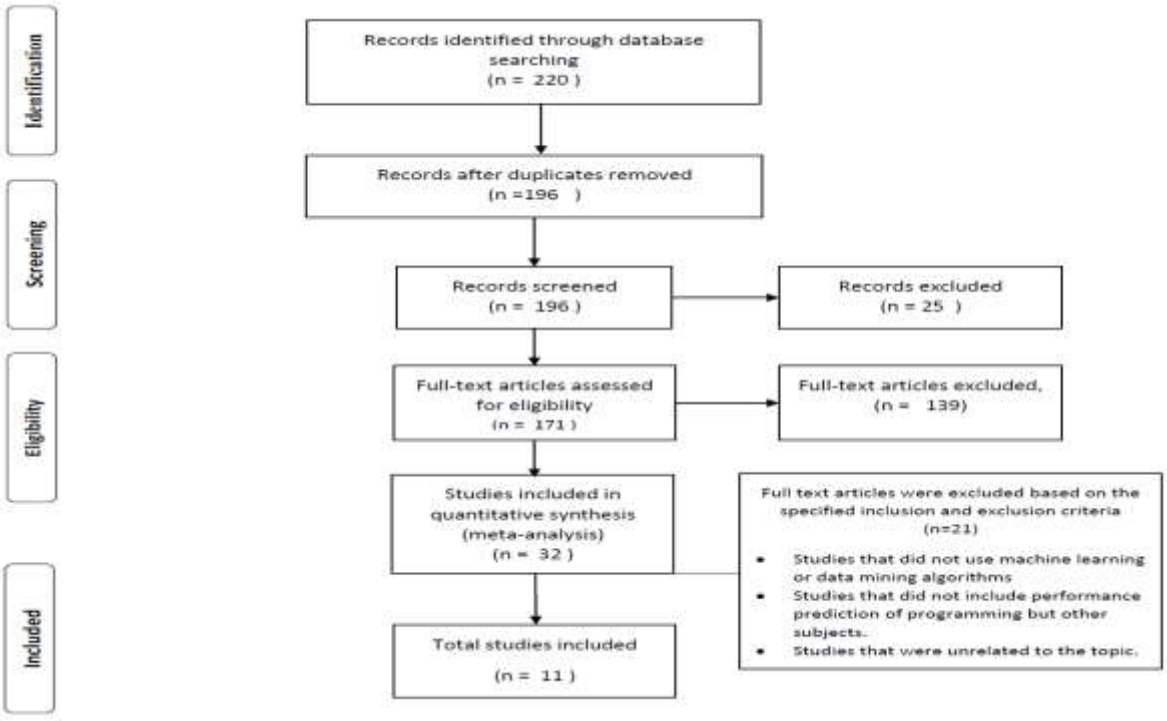


Fig. 1. Flow Diagram used for the Database Searches- PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis).

IV. Results

A. Performance of Various Algorithms

The meta-analysis comprised of eleven published studies and all eleven studies were considered in the estimation of the pooled performance of algorithms used to make the prediction. The stratification was done based on the different types of algorithms used in the extracted articles. The minimum performance of algorithm prediction was 10% and it was found in studies performed with drop out and retention. Conversely, the maximum algorithm prediction performance was found to be 36%, in a study performed with the associated student-related sub group data. The I^2 test statistic revealed high heterogeneity (I^2= 99.17%, P= <0.000). By means of the random effect analysis, the pooled performance of the algorithms was 24% (95% CI (13%, 35%). Subgroup analysis based on the types of algorithm techniques showed that the performance of the algorithm with a study using hybrid and data mining was found to be 3% (95% CI: 1%, 5%) and 20% (95% CI: 9%, 32%), respectively (Fig. 2). The midpoint and the length of each segment showed performance and a 95% CI, while the diamond shape indicated the combined performance of all studies.

B. Publication Bias

All the studies that were part of the meta-analysis were visually evaluated for publication bias using the funnel plot. Studies documented in the literature have suggested evaluating publication bias in meta-analysis to draw a reasonable conclusion about the generalizability of cumulative findings that can be affected by biases. The aim was to identify the degree to which biasness influences the study outcome to determine the validity of core findings. The funnel plot is a standard visual method for identifying publication bias. It is a scatterplot of odd log-ratio standard errors against the study effects size computed by the odd log ratio. In a funnel plot depicting a meta-analysis with no publication bias, studies will be symmetrically distributed on either side of the vertical line marking the pooled effect size if no relevant findings are missing. The funnel plot asymmetrically indicated the presence of publication bias since a higher percentage (82%) of the studies fell outside the triangular region (Fig. 3). This implies that only a smaller proportion (18%) of the studies fell inside the triangular region. In addition, the result of Egger's test revealed the presence of publication bias, P-values <0.05 (Table III). The presence of publication bias was assessed subjectively using funnel plots and objectively using the Egger's test. Each point in the funnel plots indicated a separate study and the asymmetrical distribution of studies on the plot is an indication of publication bias. First, studies' effect sizes were plotted against their standard errors and the assessment of the funnel plots revealed that in all cases the funnel plots were slightly asymmetrical (Fig. 3).

The visual examination of a funnel plot can be generally subjective to interpretation for which the Egger asymmetry method has been suggested as a complementary statistical test for bias.

The Egger test's purpose was to perform a simple linear regression to test whether the model intercept significantly differs from zero at P< 0.05. however, the funnel plots was also objectively assessed by means of Egger's weighted regression statistics. According to the symmetry assumptions, there is a publication bias in the combined (p = 0.013), pooled estimates of algorithms (Table III).

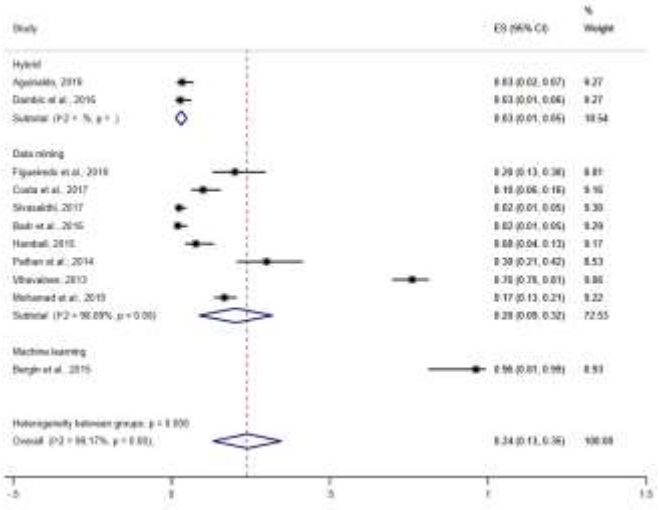


Fig. 2. The Pooled Estimates of the Performance Algorithms from Random effect Model by Type of Algorithms.

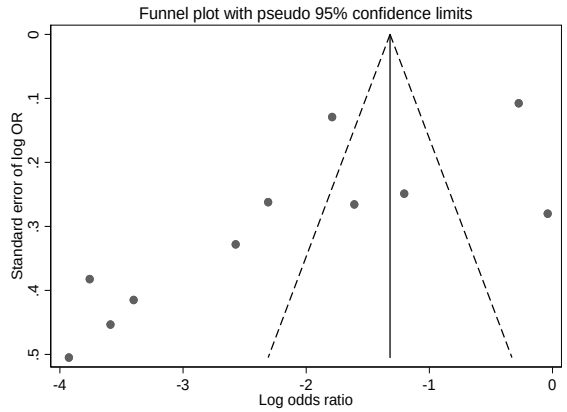


Fig. 3. Funnel Plot of the Performance of Algorithms for the Prediction.

TABLE III. EGGER'S TEST

Std-Eff	Coefficient	Std. Error	T	P>|t|	95% Confidence Interval
Slope	0.0899	0.5202	0.17	0.866	-1.0867, 1.2667
Bias	-7.4376	2.4026	-3.10	0.013	-12.8727, -2.0026

C. Sensitivity Analysis

Besides, a sensitivity test was conducted to determine the influence of each study. The outcome of the sensitivity test suggested that there was no influence on the pooled estimate of algorithm while eliminating one study at a time from the analysis. We did the sensitivity analysis of the performance of algorithms by the application of a random-effects model (Table IV). The analysis was conducted to determine the effect of each study on the pooled estimated performance of algorithms by excluding each study incrementally. The outcome of this indicated that studies that were excluded had no significant difference on the performance of algorithms.

TABLE IV. SENSITIVITY ANALYSIS OF THE INCLUDED STUDIES TO ESTIMATE THE POOLED PERFORMANCE OF ALGORITHMS

Study omitted	Performance of algorithms (95% CI)
Aguinaldo, 2019	0.281 (0.247, 0.321)
Figueiredo et al., 2019	0.272 (0.239, 0.311)
Costa et al., 2017	0.285 (0.250, 0.326)
Sivasakthi, 2017	0.287 (0.253, 0.328)
Badr et al., 2016	0.279 (0.246, 0.317)
Dambic et al., 2016	0.280 (0.246, 0.319)
Bergin et al., 2015	0.248 (0.218, 0.283)
Hambali, 2015	0.281 (0.247, 0.320)
Pathan et al., 2014	0.265 (0.232, 0.302)
Vihavainen, 2013	0.146 (0.124, 0.171)
Mohamad et al., 2010	0.314 (0.271, 0.364)

Sensitivity analysis is crucial to evaluate the robustness of combined estimates to different assumptions and inclusion criteria. The combined estimates were obtained by excluding studies judged to be at high risk of bias with those judged to be at low or moderate risk of bias [37, 38]. Hence, the presented sensitivity analysis indicated that the meta-analysis is fairly robust to the publication bias. Furthermore, the sensitivity analysis was used to assess the effects of probable violations of modelling assumptions, all of which produced alike results.

V. LIMITATIONS

The one superficial limitation of meta-analysis that has been observed in this study is the exclusion of articles that do not satisfy all the inclusion criteria. Such articles that were excluded may contain useful information. Besides, another limitation of the current study is that only the perspective of students was considered. Extending the study to capture other institutions' perspectives apart from learning institutions could have yielded more insightful findings. However, this meta-analysis study has provided valuable information regarding the most effective Educational Data Mining approaches to predict the performance of students pursuing computer programming. These limitations could be addressed in the future study because we might have missed a few relevant studies through the exclusion criteria. Further research is needed to explore the interdependencies among factors that can be utilized to predict the performance of students pursuing computer programming. In the future, we plan to explore ways to analyze missing data in related articles to cover the vital information that may have been lost because of the exclusion criteria of this study.

VI. CONCLUSION

A meta-analysis method has been used to identify and analyze factors influencing student performance, but this is the first study that applied meta-analysis to obtain the most effective Educational Data Mining approaches used to predict students' performance pursuing computer programming. Effect sizes were determined, variations and bias were determined for the included studies because of different

classifications of algorithms applied to identify students' performance pursuing computer programming. The obtained results showed that the pooled estimate of the most effective Educational Data Mining approaches used to predict students' performance pursuing computer programming was highly prevalent among participants. An attempt was made to determine the possible sources of heterogeneity by means of subgroup analysis, meta-regression, and sensitivity analysis; however, the sources of variability could not be established in all cases. The most likely reason for this colossal heterogeneity is that some of the studies were obtained from the variation among the sample size utilized in adopting the various algorithms.

REFERENCES

[1] Zaffar, M., M.A. Hashmani, and K. Savita. A study of prediction models for students enrolled in programming subjects. in 2018 4th International Conference on Computer and Information Sciences (ICCOINS). 2018. IEEE.

[2] Abdulsalam, S., et al., Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students' Performance in Computer Programming. Journal of advances in scientific research & its application (JASRA), 2015. 2: p. 79-92.

[3] Ayub, M., et al., Modelling students' activities in programming subjects through educational data mining. Global Journal of Engineering Education, 2017. 19(3): p. 249-255.

[4] Cavoukian, A., Data mining: Staking a claim on your privacy. Information and privacy, Commissioner Ontario. 1998.

[5] Pathan, A.A., et al. Educational data mining: A mining model for developing students' programming skills. in The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014). 2014.

[6] Ventura, S., C. Romero, and C. Hervás. Analyzing rule evaluation measures with educational datasets: A framework to help the teacher. in Educational Data Mining 2008. 2008.

[7] Guo, B., et al. Predicting Students Performance in Educational Data Mining. in 2015 International Symposium on Educational Technology (ISET). 2015.

[8] Farid, D.M., et al., An adaptive ensemble classifier for mining concept drifting data streams. Expert Systems with Applications, 2013. 40(15): p. 5895-5906.

[9] Farid, D.M., et al., Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert systems with applications, 2014. 41(4): p. 1937-1946.

[10] Buniyamin, N., U.b. Mat, and P.M. Arshad. Educational data mining for prediction and classification of engineering students achievement. in 2015 IEEE 7th International Conference on Engineering Education (ICEED). 2015

[11] Mohamad, S.K. and Z. Tasir. Pattern of Reflection in Learning for Predicting Students' Performance. in 2014 International Conference on Teaching and Learning in Computing and Engineering. 2014.

[12] Romero, C. and S. Ventura, Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2010. 40(6): p. 601-618.

[13] Manjarres, A.V., L.G.M. Sandoval, and M.S. Suárez, Data mining techniques applied in educational environments: Literature review. Digital Education Review, 2018(33): p. 235-266.

[14] Sikder, M.F., M.J. Uddin, and S. Halder. Predicting students yearly performance using neural network: A case study of BSMRSTU. in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV). 2016. IEEE.

[15] Pandey, M. and S. Taruna, A comparative study of ensemble methods for students' performance modeling. International Journal of Computer Applications, 2014. 103(8).

[16] Kaur, G. and W. Singh, Prediction of student performance using weka tool. An International Journal of Engineering Sciences, 2016. 17: p. 8-16.

[17] Barlow-Jones, G. and D. van der Westhuizen. Pre-entry attributes thought to influence the performance of students in computer programming. in Annual Conference of the Southern African Computer Lecturers' Association. 2017. Springer.

[18] Hostetler, T.R., Predicting student success in an introductory programming course. SIGCSE bulletin, 1983. 15(3): p. 40-43.

[19] Sivasakthi, M. Classification and prediction based data mining algorithms to predict students' introductory programming performance. in 2017 International Conference on Inventive Computing and Informatics (ICICI). 2017.

[20] Đambić, G., M. Krajcar, and D. Bele, Machine learning model for early detection of higher education students that need additional attention in introductory programming courses. International Journal of Digital Technology & Economy, 2016. 1(1): p. 1-11.

[21] Costa, E.B., et al., Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in human behavior, 2017. 73: p. 247-256.

[22] Figueiredo, J., N. Lopes, and F.J. García-Peñalvo, Predicting Student Failure in an Introductory Programming Course with Multiple Back-Propagation, in Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality. 2019, Association for Computing Machinery: León, Spain. p. 44–49.

[23] Vihavainen, A. Predicting Students' Performance in an Introductory Programming Course Using Data from Students' Own Programming Process. in 2013 IEEE 13th International Conference on Advanced Learning Technologies. 2013.

[24] Bergin, S., et al., Using machine learning techniques to predict introductory programming performance. International Journal of Computer Science and Software Engineering (IJCSSE), 2015. 4(12): p. 323-328.

[25] Aguinaldo, B.E. 21st Century learning skills predictive model using PART algorithm. in ACM International Conference Proceeding Series. 2019.

[26] Akinola, O., B. Akinkunmi, and T. Alo, A data mining model for predicting computer programming proficiency of computer science undergraduate students. 2012.

[27] Mohamad Farhan Mohamad, M., et al. Mining the student programming performance using rough set. in 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering. 2010.

[28] Badr, G., et al., Predicting students' performance in university courses: a case study and tool in KSU mathematics department. Procedia Computer Science, 2016. 82: p. 80-89.

[29] Moher, D., et al., Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic reviews, 2015. 4(1): p. 1.

[30] Tam, W.W., et al., Perception of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement of authors publishing reviews in nursing journals: a cross-sectional online survey. BMJ open, 2019. 9(4): p. e026271.

[31] Zhang, P. and B. Liu, Differentiation among Glioblastomas, Primary Cerebral Lymphomas, and Solitary Brain Metastases Using Diffusion-Weighted Imaging and Diffusion Tensor Imaging: A PRISMA-Compliant Meta-analysis. ACS Chemical Neuroscience, 2020. 11(3): p. 477-483.

[32] Munn, Z., et al., The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence. International journal of health policy and management, 2014. 3(3): p. 123.

[33] Fabrizi, F., et al., HBV infection is a risk factor for chronic kidney disease: Systematic review and meta-analysis. Revista Clinica Espanola, 2020.

[34] Fahmy, T. and A. Bellétoile, Algorithm 983: Fast Computation of the Non-Asymptotic Cochran's Q Statistic for Heterogeneity Detection. ACM Transactions on Mathematical Software (TOMS), 2017. 44(2): p. 1-12.

[35] Begg, C.B. and M. Mazumdar, Operating characteristics of a rank correlation test for publication bias. Biometrics, 1994: p. 1088-1101.

[36] Egger, M., et al., Bias in meta-analysis detected by a simple, graphical test. Bmj, 1997. 315(7109): p. 629-634.

[37] Kristiansen, M., Dhami, S., Netuveli, G., Halken, S., Muraro, A., Roberts, G., Larenas-Linnemann, D., Calderón, M.A., Penagos, M., Du Toit, G. and Ansotegui, I.J., 2017. Allergen immunotherapy for the prevention of allergy: a systematic review and meta-analysis. Pediatric Allergy and Immunology, 28(1), pp.18-29.

[38] Mathur, M.B. and VanderWeele, T.J., 2020. Sensitivity analysis for publication bias in meta-analyses. Journal of the Royal Statistical Society: Series C (Applied Statistics), 69(5), pp.1091-1119.

Adaptive Congestion Window Algorithm for the Internet of Things Enabled Networks

Ramadevi Chappala[1], Ch.Anuradha[2], Dr.P. Sri Ram Chandra Murthy[3]
Research Scholar, Department of CSE, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India[1]
Assistant Professor, Department of CSE, V.R.K. Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India[2]
Assistant Professor, Department of CSE, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India[3]

***Abstract*—Heterogeneous constrained computing resources in the Internet of Things (IoT) are communicated, collected, and share information from the environment using sensors and other high-speed technologies which generate tremendous traffic and lead to congestion in the Internet of Things (IoT) networks. This paper proposes an Adaptive Congestion Window (ACW) algorithm for the Internet of Things. This algorithm is adapted to the traffic changes in the network. The main objective of this paper is to increase the packet delivery ratio and reduce delay while enhancing the throughput which can be attained by avoiding congestion. Therefore, in the proposed algorithm, the congestion window size is depending on the transmission rate of the source node, the available bandwidth of the path, and the receiving rate of the destination node. The congestion window size is altered when the link on the path needs to be shared/released with/by other paths of different transmission in the network. The proposed algorithm, ACW is simulated, evaluated in terms of packet delivery ratio, throughput, and delay. The performance of the proposed algorithm, ACW is compared with IoT Congestion Congrol Algorithm (IoT-CCA) and Improved Stream Control Transmission Protocol (IMP-SCTP) and proved to be better by 27.4%, 11.8%, and 33.7% than IoT-CCA and 44.1%, 22.6%, and 50% than IMP-SCTP concerning packet delivery ratio, throughput, and delay respectively. The variation in congestion window size with time is also projected.**

Keywords*—Congestion window; internet of things; packet delivery ratio; throughput

I. INTRODUCTION

Internet of Things (IoT) is emerging technologies where it connects and shares information globally from kitchen set to cars to industrial tools also target to interrelate and incorporate the physical world and information technology. IoT is the source for every industry to stand in the market and every industry realize that IoT is the key for development in the industries using various IoT technologies like Industrial Internet of Things (IIoT), Internet of Logistics Things (IoLT), Internet of Retail Things (IoRT), Internet of Workforce Management (IoWM) and Internet of Medical Things (IoMT). With these technologies, IoT has an extensive application including smart cities, smart homes, smart communities, utilities and appliances, intelligent transportation, industrial production, E-health, military and environmental monitoring. To acquire the information from various devices of the world, IoT uses capabilities of computing, communication, and perception by utilizing actuators and sensors. These sensors, actuators, or any physical device can be smart objects with the ability to sense, collect data from the environment, communicate and interact with these physical objects. These objects are smart because of their intelligent behaviour for their connection, communication using a wireless protocol. Congestion can be presented in both wired and wireless networks in an IoT environment. And the intelligent environment is created with the interconnection of IoT devices and these IoT devices generate tremendous traffic. Internet of Things (IoT) is attaining enormous study consideration because of the necessity of assimilation of various kinds of networks. Connecting more and more devices to provide various services which share information among them is the main objective of IoT. Presently, every single individual who is associated with the web is utilizing distinctive kinds of specialized gadgets. For two decades, IoT is increasing its popularity and huge work is being carried out by different analysts and business or investors. The objective of IoT is to construct our everyday social lives to be simple [1].

To control the congestion in both wired and wireless networks according to the transmission rate, Transmission Control Protocol (TCP) is the most reliable, connection-oriented transport layer protocol. Network bandwidth and delay modified according to network conditions. There are various application protocols like Advanced Message Queuing Protocol (AMQP) [2], Extensible Messaging Presence Protocol (XMPP) [3], XMPP Representaional State Transfer Hyper Text Transfer Protocol (XMPP RESTful HTTP) and Message Queuing Telemetry Transport (MQTT) [4] are the application protocols presented in Fig. 1, supported by TCP in IoT environment. Another application protocol running over the connectionless, unreliable transport protocol UDP is Constrained Application Protocol (CoAP) is used to provide communication among various gadgets in IoT networks. The open-source community-developed XMPP based on Extensible Markup Language (XML) for instant messaging in a real-time environment which supports the process to process communication among devices in IoT network. MQTT is another application protocol introduced by Arlen Nipper and Dr.Andy Stanford-Clark which is a lightweight protocol for machine-to-machine communication with different services like congestion-controlled, reliable, process2process, and connection-oriented facilities. Among heterogeneous devices in an IoT environment, AMQP provides data transfer services.

Wireless sensor networks (WSN) have been presumed a critical part of communication because of their extraordinary highlights (e.g., versatility and simplicity of association) that make them a significant transporter of information across networks [5-7]. The unique fundamental driver of the lessening in the lifetime of nodes in WSN and reduction in node's energy is due to lack of congestion control [8, 9]. This lessening prompts numerous different issues, for example, delay, loss of packets, and transmission capacity deprivation [10]. Different applications like query-focused, uninterrupted sensing, hybrid applications, event-based, etc are influenced by congestion control [11].

The rest of the paper is structured as follows: the outline of the related work is presented in Section 2. The proposed work is presented in Section 3. Results are discussed in Section 4 and finally, Section 5 concludes the paper.

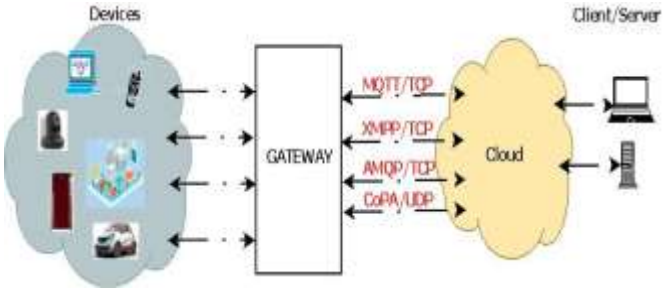


Fig. 1. IoT Application Protocol.

II. Related Work

With the introduction of mobile networks and smartphone technology, services like Internet of Vehicles, Internet of Things, Device to Device communication, and mobile networks undergoes many modifications and also most of the real-time implementations request for maximum throughput and least end-to-end delay which leads to congestion in the network. Wired backbone connections are used to integrate the mobile networks, with these changes the network and transport layers are affected a lot. To control the congestion, TCP accomplished various window-based flow control techniques uses CWND (congestion window) and SSTHRESH (slow-start threshold) which are two state variables. These two variables are used to control the rate of transmission in the network. The objective of CWND is to allow a sender to send data not more than the maximum capacity of the network in any condition and it automatically adjusts to the present network status. Threshold value is provided by slow start threshold variable to control congestion and these two variables are modified by TCP variants. Due to packet loss, the TCP sender identifies that congestion takes place in the network either by duplicate acknowledgment or timeout mechanism. CWND and SSTRHRESHOLD modified by the sender to TCP protocol which plays a key role in IoT development. Different congestion control mechanisms are supported by TCP variants in the IoT environment. In delay networks usage of bandwidth corresponding to the received acknowledgment is a great challenge for researchers. Therefore, in the proposed algorithm, the congestion window size is depending on the transmission rate of the source node, the available bandwidth of the path, and the receiving rate of the destination node. The congestion window size is altered when the link on the path needs to be shared/released with/by other paths of different transmission in the network.

As IoT applications increased day by day and the gadgets which are connected and communicated rose the huge amount of traffic in IoT networks. IoT consists of WSN along with the software as a separate layer that is installed among the computational devices over the cloud. Zigbee is the most familiar WSN protocol based on which IoT is implemented. This paper gives a review of different congestion control procedures utilized at the transport layer. Accessible congestion control procedures, their pros and cons, and prevailing issues with TCP in IoT are also incorporated in [12].

A decision tree (DT) is an AI model that makes more acceptable congestion control in 5G IoT networks. To decide an ideal parametric setting in a 5G network this framework was performed on a training dataset. To improve the behavior of the congestion control method, a dataset was employed to construct the AI model. A decision tree can be used with various capabilities, especially in estimation and grouping. To understand the estimation procedure by any client the DT method will give results [13].

An AI model was used to enhance congestion control, and the methodology delivered an optimistic outcome for the practical and uncertain evaluations of route protocols [14, 15]. Sangeetha et al. [16] suggested a decrease in energy and information misfortune due to congestion over the network. Essentially, the sensor node topology is adjusted intermittently at periodic time interim and node level to upgrade the power utilization of sensor nodes, the intervention, and give an "energy-efficient congestion aware routing procedure for WSNs—specifically, survivable path routing (SPR)".

Singh et al. [17] introduced a new congestion control method towards WSNs but the traditional procedures have more power utilization with more intricacy also got the ideal rate by retransmission with congestion control utilizing the basic Poisson procedure. The routing procedure to choose the ideal path projected by Shelke et al. [18], because of opportunistic hypothesis and by coordinating reasonable rest planning components to diminish congestion in the organization, builds singular node life, the whole organization lifetime, and diminishes division in the organization. Godoy et al. [19] researched and examined the conditions that lead to congestion of the correspondence channel dependent on node setup boundaries: transmission periods, the rate at which packets are generated, and transmitter yield power level.

High-speed TCP is proposed by Floyd in [20] after identifying the effectiveness issue in a high-speed network. This procedure utilizes α to avoid congestion and β as a reduction factor for the duration of trivial loss discovery. The drawbacks of this procedure are dealt with in Scalable TCP which is presented in [21]. The "Multiplicative increase and Multiplicative Decrease" is used in Scalable TCP. The inter-fairness issue is the drawback eliminated in HTCP [22]. A lapse period is introduced before the most recent congestion occurrence in HTCP. BIC-TCP is proposed in [23] which is improved in TCP-CUBIC [24]. For estimating the size of the

congestion window and RTT fairness, TCP-CUBIC uses the CUBIC function which is not depending on RTT. It makes use of packet loss as a congestion indicator but does not fully utilize the resources as it is not difficult than H-TCP and HS-TCP and it is used in the Linux kernel. TCP-CUBIC is enhanced in CUBIC-FIT [25].

A novel congestion control strategy is presented in [26] to adjust the transmission rate rapidly at whatever point the accessible transfer speed besides various delay. The suggested approach keeps up a consistent situation to decrease packet loss along with maximizing throughput. And also present versatile procedures to keep up reasonableness with broadly installed TCP Cubic.

A congestion control procedure called TCP Vegas based on delay is developed in [27]. It changes the cwnd as per the distinction between the estimated and the real rate. This procedure increases packet delivery ratio, yet it experiences low data transmission usage in fast networks. This genuine unfairness issue of TCP Vegas is perceived and suggested a new TCP variation known as Vegas^{+} in [28]. TCP Westwood [29] approximates the accessible transmission capacity dependent on the rate at which acknowledgments are received.

Cheng Ding et al. [30] have suggested a mechanism to allocate cluster-head nodes uniformly, node clustering approach utilizes nodes with maximum traffic and more residual energy depending on energy consumption optimization and energy balancing. Depending on load balance the authors utilize a data forwarding approach to choose suitable routes for various services for delay and service priority specifications to differentiate various services in the network environment.

Al-Janabi et al. [31] have proposed a systematic algorithm depending on load adjustment for IoT-based SDN known as clustering algorithm makes use of storage units and data canters situated on the cloud as cloud resources to evaluate load-balanced PSO clustering algorithm. To build a clustering table, the PSO clustering algorithm utilizes transmission cost, load balancing, and other energy components in the SDN controller where the cluster table utilizes cluster members and cluster heads information of cluster.

Hussien Saleh Altwassi et al. [32] have instigated a well-structured load balancing protocol to improve the life span of the network and to estimate the transmission quality with the metrics power consumption and packet delivery ratio to minimize the congestion in RPL (IPV6 Routing Protocol for Low Power Lossy networks) networks.

Arfath Azeez et al. [33] provide various services to access the server's information and publish/subscribe mechanism to clients which are connected to servers using MQTT cloud-based protocol. The suggested mechanisms incorporate the Application Delivery Controller (ADC) which upgrades and controls the way how the client communicates to the data center to cache/read information or server for refining among the gadgets and the data center. And also takes responsibility to change the route for the entire data to different datacenters when it falls to handle the request or crashes as well as transfer data to the data center having less or nothing when load rises in the network.

Santiago et al. [34] have proposed an energy balancing algorithm that diminishes the energy consumption by selecting the better parent node with the event rate. From the results, it is proved to be that the life span of the parent node and the network is raised and also minimizes the energy consumption.

The main objective of this paper is to increase the packet delivery ratio and reduce delay while enhancing the throughput which can be attained by avoiding congestion. Therefore, in the proposed algorithm, the congestion window size is depending on the transmission rate of the source node, the available bandwidth of the path, and the receiving rate of the destination node. The congestion window size is altered when the link on the path needs to be shared/released with/by other paths of different transmission in the network.

III. Adaptive Congestion Window Algorithm for Internet of Things

The devices in IoT are heterogeneous and also keep increasing and hence more and more communication will be carried out at a time. The IoT devices continuously sense and transmit information. The IoT devices might communicate among themselves or with the cloud or with any other network like hospitals, etc.

Let us consider the devices in the IoT network as nodes and the communication among the nodes/devices will be carried out using wireless links and are represented with the dashed line. The sample scenario is shown in Fig. 2. The communication among the nodes is shown using black color dashed line. The communication between the devices and the network1/network2 is shown using a pink color dash and a single dotted line. The communication with the cloud is shown using a brown color dash and two dotted lines.

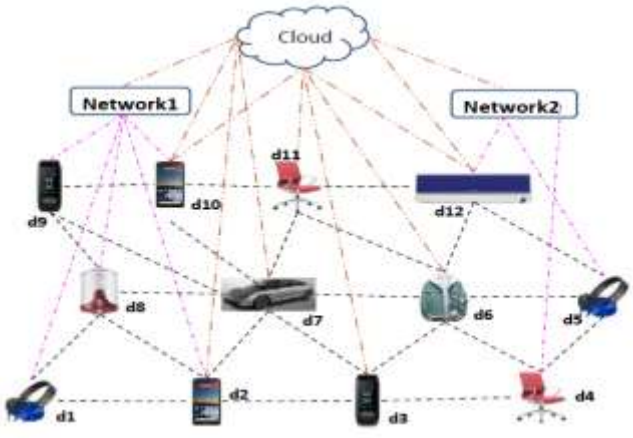


Fig. 2. Sample Scenario of IoT Network.

There might be 'n' communications taking place at the time 't' which might be increasing or decreasing at the time 't+1'. Hence, the available bandwidth of a particular link will be changing from time to time. These changes in the available bandwidth of the links may lead to congestion in the network if the source node does not adapt its transmission rate / cwnd (congestion window) accordingly. Therefore, in this paper, the cwnd depends on the transmission rate with which the sender can transmit also the available bandwidth of the path and speed with which the receiver can receive the information.

Let us consider:

The transmission rate with which the sender/source node can transmit the information – R_{Tx}

The receiving rate with which the receiver/destination node can receive the information– R_{Rx}

The available bandwidth of the path = min (available bandwidth of the links in the corresponding path) –BW_{avl} -- eqation (1)

Then the congestion window, cwnd = min (R_{Tx}, R_{Rx}, BW_{avl}) -- equation (2)

The selection of the congestion window in this manner will reduce the number of packet drops as there would be a sufficient amount of bandwidth to transmit the information. The packet drops will not be only due to congestion, it might be because of any other network issue like node breakdown, link breakage, etc. This is handled by changing the path of the transmission. So, the cwnd size is modified only there is a change in the available bandwidth.

Initially, the congestion window size is determined using eq. 1 between a set of sender s1 and receiver r1. Later, during the transmission between s1 and r1, if there are any path changes or network changes, then there is a chance of change in the available bandwidth as the common link need to share its available bandwidth amid both the paths in which it is involved. In this regard, the source of the link is responsible to inform the source node of the path to change its cwnd size accordingly.

Hence, the congestion window size is recomputed. The available bandwidth might also change when any link of the selected path is being shared or stopped sharing by any other path of other transmissions as more number of transmissions among various nodes can be carried out synchronously. Periodically, the packet delivery ratio is computed. Whenever the packet delivery ratio is below the threshold value, the congestion window size is reduced. Algorithm 1 is executed when any node is ready for transmission. Algorithm 2 is executed when any node stops or completes its transmission of all the information.

Algorithm 1

Input:

Number of nodes – n
A transmission rate of nodes – R_{Txi}
Receiving rate of nodes – R_{Rxi}

Output:

cwnd size of the path, p_j

Begin

1. Determine the path, p_i between the senderT_{xi} and receiverR_{xi}
2. For all paths p_j
   a. If any link of the path, p_i is common to the path, p_j then
      i. The available bandwidth is shared among paths, p_i, and p_j
      ii. The source node of the common link informs about the change in available bandwidth to the source node of the path, p_i
      iii. The source node of the path, p_i updates the cwnd size of the path, p_i
3. Determine the available bandwidth of all the links of the path determined in 1.a
4. Determine the available bandwidth, BW_{avli} of the path, p_i using the eq. (1)
5. Determine the cwnd size using eq. (2): cwnd = min(R_{Txi}, R_{Rxi}, BW_{avli})
6. Start the transmission
7. After every transmission
   a. If (PDR_i<PDR_{th}) then
      i. Reduce cwnd size by 10%
8. If any more packets to be transmitted than
   a. Goto 6.

End

Algorithm 2

Input:

p_i – Transmission stops along this path

Begin

1. For all paths p_j
   a. If any link of the path, p_i is common to a path, p_j then
      i. The bandwidth of this common link is released by path, p_i
      ii. The source node of the common link informs about the change in available bandwidth to the source node of the path, p_j
      iii. The source node of the path, p_j updates the cwnd size of path, p_j

End

Consider the network shown in Fig. 2.

Assume that,

Node d1 is transmitting the information to node d12 using the path, d1 → d8 → d7 → d6 →d12.

The transmission rate with which the Node d1 can transmit the information – 80Mbps

The rate at which the node d12 can receive the information – 65Mbps

The available bandwidth of this path –72Mbps

Hence, the cwnd size is set to 65 Mbps as min(80,65,72) = 65Mbps.

At this instant, assume that node d2 communicates with d9 using the path, d2 → d7 → d9. As there are no common links, there will no change in the cwnd size of node d1.

After some time, assume that node d4 starts its communication with d9 using the path, d4 → d6 → d7 → d9.

It can be observed that the link d6 → d7 is common in both the communications between d1 – d12 andd4 – d9. Also, the link d7 → d9 is common in both the communications between d2 – d9 and d4 – d9. When node d4 starts its communication, the bandwidth of the links d6 → d7 and d7 → d9 need to be shared among both the paths. Therefore, d6 informs d1 about the change in the available bandwidth of the link, d6 → d7, and d7 informs d2 about the change in the available bandwidth of the link, d7 → d9. Hence, the cwnd size of source nodes d1, d2, and d4 is determined accordingly. Later, when the node d2 completes its communication, the cwnd size of d4 needs to be updated but d1 needs not to be updated.

IV. Results and Discussion

The proposed algorithm, the Adaptive Congestion Window (ACW) algorithm is simulated using NS-2. The simulation is executed for 150s. The topology used for the simulation purpose includes 50 nodes and cloud environments. All the nodes in the network are capable of transmitting, receiving, and forwarding the packets. The bandwidth and propagation delay of links between nodes is different. The traffic type used for simulation purposes includes both CBR and VBR. Both types of traffic are included as the information might be in the form of text, images, audio, or video. The parameters used to evaluate the performance of the proposed ACW algorithm are throughput, packet delivery ratio, and delay for time and congestion window size. The proposed algorithm, ACW proved to be performing better when compared with IMP-SCTP [10] and IoT-CCA[23] in terms of throughput, packet delivery ratio, and delay.

The performance of the proposed algorithm, ACW in terms of packet delivery ratio in comparison with IMP-SCTP and IoT-CCA is shown in Fig. 3 and proved to be performing better. The results are shown at varying times. It can be observed that the packet delivery ratio decreases as time increases. The packet delivery ratio is better when compared to IMP-SCTP and IoT-CCA as the congestion window size is based on the transmission rate of the sender, the available bandwidth of the path, receiving rate of the destination node. As the congestion window size is less than or equal to the amount of bandwidth available, there are fewer chances of packets being dropped. There are some packets dropped because of unavoidable or unexpected issues in the network like link breakage or node breakdown. The performance of ACW is 27.4% better than IoT-CCA and 44.1% better than IMP-SCTP in the case of packet delivery ratio.

The performance of the proposed algorithm, ACW is compared with IMP-SCTP and IoT-CCA in terms of throughput with varying time is shown in Fig. 4. It can be observed that ACW performs better than IMP-SCTP and IoT-CCA. The throughput decreases as time increases. The throughput is dependent on the packet delivery ratio. As there is an improvement in the packet delivery ratio, throughput also enhances. The performance of ACW is 11.8% better than IoT-CCA and 22.6% better than IMP-SCTP in the case of throughput.

The efficiency of the proposed algorithm, ACW is compared with IMP-SCTP and IoT-CCA in terms of delay with varying time is shown in Fig. 5. It can be observed that ACW performs better than IMP-SCTP and IoT-CCA. The delay increases as time increases. The delay occurs due to the minimum propagation delay at nodes in the path. When the number of nodes decreases, then the delay can be further reduced. As the minimum available bandwidth is considered to be the available bandwidth of the path which is one of the parameters to determine the congestion window size, mostly the packets need not be buffered at intermediate nodes. This leads to the reduction of delay in ACW when compared to IMP-SCTP and IoT-CCA. The enhancement of performance of ACW in terms of delay in comparison with IoT-CCA is 33.7% and IMP-SCTP is 50%.

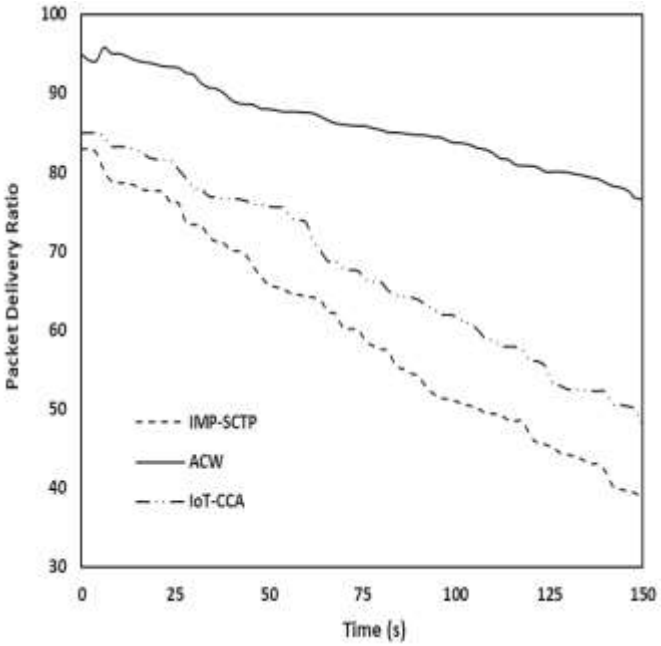


Fig. 3. Packet Delivery Ratio vs Time.

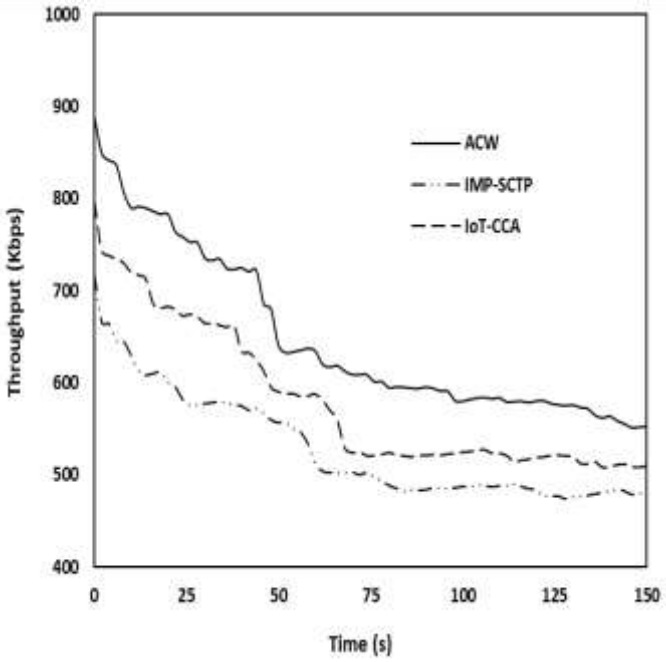


Fig. 4. Throughput vs Time.

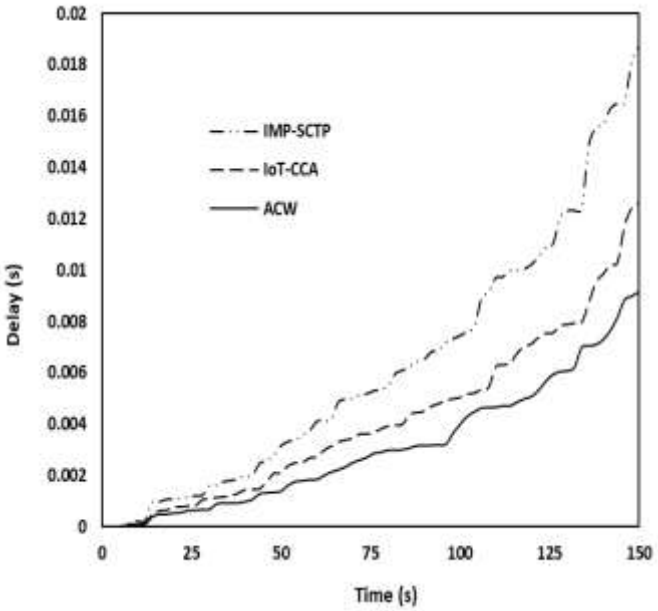


Fig. 5. Delay vs Time.

Fig. 6 shows how congestion window size varies with time for the proposed algorithm, ACW. This graph is plotted for a particular pair of source and destination nodes. The size of the congestion window is either increased or decreased depending on the links in the path and also the links being shared with other paths. The congestion window size is decreased if one of the available links in the path is needed to be shared with another path using which another transmission is about to start. The size of the congestion window is raised if any of the links of the path being shared are released by another path.

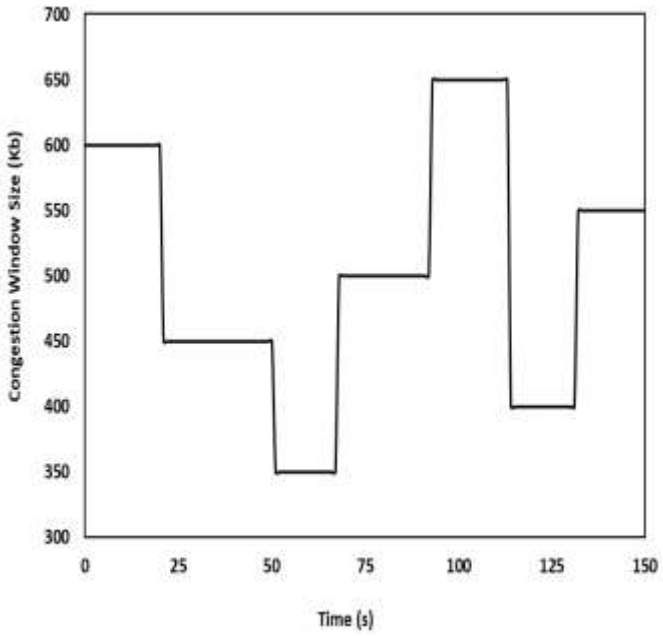


Fig. 6. Congestion Window Size vs Time.

V. Conclusion

For the Internet of Things (IoT), this paper proposes an adaptive congestion window algorithm. Congestion window size (cwnd) of the proposed algorithm, ACW is dependent on the transmission rate of the source node, the available bandwidth of the path, receiving rate of the destination node. The congestion window size of a particular path is increased or decreased with the release/sharing of the available bandwidth of one of the links in the path. The results of the proposed algorithm, ACW are simulated and evaluated in terms of packet delivery ratio, throughput, and delay. The variation of the congestion window size for time is also shown. The performance of the proposed algorithm is compared against IMP-SCTP and IoT-CCA and proved to be better. And in future, this algorithm can be tested against more parameters by including the priority of the nodes to prove that the future results would be better than the proposed work and enhance the performance of network utilization in IoT sensor networks.

Acknowledgment

The author is grateful to Acharaya Nagarjuna University for providing facilities to carry out research work. Also, I would like to acknowledge the Dean of Acharya Nagarjuna University for the comments provided while preparing this work

References

[1] Vermesan, O., Friess, P., Guillemin, P., Gusmeroli, S., Sundmaeker, H., Bassi, A., & Doody, P. (2011). Internet of things strategic research roadmap. January, pp. 9-52.

[2] Saint-Andre,P.(2011)"Extensible messaging and presence protocol(XMPP): ore," IETF RFC 6120, https://tools.ietf.org/html/rfc6120 Accessed 12-June-2019.[15]J.Postel,(1981). Transmission Control Protocol. https://tools.ietf.org/html/rfc793. Accessed 12-June-2019.

[3] Saint-Andre,P.(2011).Extensible Messaging and Presence Protocol (XMPP):Core.https://tools.ietf.org/html/rfc6120. Accessed 10 December 2018.

[4] MQTT Version3.1.1 (2015). http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/errata01/os/mqtt-v3.1.1-errata01-os-complete.doc. Accessed 12- December - 2018.

[5] Najm, A.,Ismail, M. Rahem, T. Al Razak, A. (2014). Wireless implementation selection in the higher institution learning environment. J. Theor. Appl. Inf. Technol, 67(2),pp.477-484.

[6] Ismail, M.,Najm, I.A., and Balfaqih,M.(2017).Topology sense and graph-based TSG: efficient wireless ad hoc routing protocol for WANET. Telecommunication Systems, 65(4),pp.739-754.

[7] Aalsalem, M.Y.,Khan, W.Z.,Gharibi, W.,Khan, M.K., andArshad, Q. (2018). Wireless Sensor Networks in oil and gas industry: Recent advances taxonomy requirements and open challenges. Journal of Network and Computer Applications, 113,pp.87-97.

[8] Sunny, A., Panchal, S., Vidhani, N., Krishnasamy, S., Anand, S. V. R., Hegde, M., & Kumar, A. (2017). A generic controller for managing TCP transfers in IEEE 802.11 infrastructure WLANs. Journal of Network and Computer Applications, 93, pp.13-26.

[9] Flora, D.J., Kavitha,V. &Muthuselvi, M. (2011). A survey on congestion control techniques in wireless sensor networks. International Conference on Emerging Trends in Electrical and Computer Technology,1146-1149. IEEE.

[10] Karunakaran,S.,and Thangaraj,P.(2010).A cluster based congestion control protocol for mobile ad hoc networks. International Journal of Information Technology and Knowledge Management, 2(2), pp.471-474.

[11] Kafi, M. A., Djenouri, D., Ben-Othman, J., and Badache, N. (2014). Congestion control protocols in wireless sensor networks: A survey. IEEE communications surveys and tutorials, 16 (3), pp.1369-1390.

[12] Mishra, N., Verma, L. P., Srivastava, P. K., and Gupta, A. (2018). An analysis of IoT congestion control policies. Procedia computer science, 132,pp. 444-450.

[13] Najm, I. A.,Hamoud, A. K., Lloret, J., and Bosch, I. (2019). Machine learning prediction approach to enhance congestion control in 5G IoT environment. Electronics, 8(6),pp. 607-614.

[14] Mirza, M., Sommers, J., Barford, P., and Zhu, X. (2007). A machine learning approach to TCP throughput prediction. ACM SIGMETRICS Performance Evaluation Review, 35(1),pp. 97-108.

[15] Kong, Y., Zang, H., and Ma, X. (2018). Improving tcp congestion control with machine intelligence. In Proceedings of the 2018 Workshop on Network Meets AI & ML pp.60-66.

[16] Sangeetha, G., Vijayalakshmi, M., Ganapathy, S., and Kannan, A. (2018). A heuristic path search for congestion control in WSN. In Industry Interactive Innovations in Science, Engineering and Technology pp.485-495 Springer, Singapore.

[17] Singh, K., Singh, K., and Aziz, A.(2018). Congestion control in wireless sensor networks by hybrid multi-objective optimization algorithm. Computer Networks, 138,pp. 90-107.

[18] Shelke, M., Malhotra, A., and Mahalle, P. N. (2018). Congestion-aware opportunistic routing protocol in wireless sensor networks. In Smart computing and informatics , pp.63-72. Springer, Singapore.

[19] Godoy, P. D., Cayssials, R. L., and Garino, C. G. G. (2018). Communication channel occupation and congestion in wireless sensor networks. Computers & Electrical Engineering, 72, pp.846-858.

[20] Floyd,S. (2003). “High Speed TCP for large congestion windows”. https://tools.ietf.org/html/rfc3649.

[21] Kelly, T. (2003). Scalable TCP: Improving performance in high speed wide area networks. ACM SIGCOMM computer communication Review, 33(2),pp.83- 91.

[22] D.Leith,RShorten(2007).“H-TCP:TCP congestion control for high bandwidth-delay product paths”. https://tools.ietf.org/html/draft-leith-tcp-htcp-03.

[23] Xu, L., Harfoush, K., and Rhee, I. (2004). Binary increase congestion control (BIC) for fast long-distance networks. In IEEE Infocom Vol. 4, pp.2514-2524. IEEE.

[24] Ha, S., Rhee, I., and Xu, L. (2008). Cubic: a new TCP-friendly high-speed TCP variant. ACM SIGOPS operating systems review, 42(5), pp.64-74.

[25] Wang, J., Wen, J., Han, Y., Zhang, J., Li, C., and Xiong, Z. (2013). CUBIC-FIT: A high performance and tcp CUBIC friendly congestion control algorithm. IEEE Communications Letters, 17(8), pp.1664-1667.

[26] Verma, L. P., and Kumar, M. (2020). An IoT based Congestion Control Algorithm. Internet of Things, 9, 100157.

[27] Brakmo, L. S., and Peterson, L. L. (1995). TCP Vegas: End to end congestion avoidance on a global Internet. IEEE Journal on selected Areas in communications, 13(8),pp.1465-1480.

[28] Hasegawa, G., Kurata, K., and Murata, M. (2000). Analysis and improvement of fairness between TCP Reno and Vegas for deployment of TCP Vegas to the Internet. In Proceedings 2000 International Conference on Network Protocols , pp.177-186. IEEE.

[29] Mascolo, S., Casetti, C., Gerla, M., Sanadidi, M. Y., and Wang, R. (2001). TCP westwood: Bandwidth estimation for enhanced transport over wireless links. In Proceedings of the 7th annual international conference on Mobile computing and networking ,pp.287-297.

[30] Ding, C., Xu, S., Chen, X., Zhou, G., Zheng, P., & Li, Y. (2019). A Delay and Load-balancing based Hierarchical Route Planning Method for Transmission Line IoT Sensing and Monitoring applications. In 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), pp.207-215. IEEE.

[31] Al-Janabi, T. A., and Al-Raweshidy, H. S. (2017). Optimised clustering algorithm-based centralised architecture for load balancing in IoT network. In 2017 International Symposium on Wireless Communication Systems, 269-274. IEEE.

[32] Altwassi, H. S., Pervez, Z., Dahal, K., and Ghaleb, B. (2018). The rpl load balancing in iot network with burst traffic scenarios. In 2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA) pp. 1-7. IEEE.

[33] Md.ArfathAzeez,D.T. Supreeth Rao and ChinmayKini (2019).Load Balancing and Crash Management in IoTEnvironment,International Research Journal of Engineering and Technology (IRJET), Vol-6, No-3, pp.46-50.

[34] Santiago, S., Kumar, A. D. V., and Arockiam, L. (2018). EALBA: energy aware load balancing algorithm for IoT networks. In Proceedings of the 2018 International Conference on Mechatronic Systems and Robots pp. 46-50.

AMBA: Adaptive Monarch Butterfly Algorithm based Information of Transfer Scheduling in Cloud for Big Information Application

D. Sugumaran[1]
Associate Professor, Department of Information Technology, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology
Chennai,Tamilnadu,India

C. R. Bharathi[2]
Associate Professor, Department of Electronics and Communications, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology
Chennai,Tamilnadu,India

***Abstract*—In present days cloud computing is most famous innovation and has a great deal of research potential in different zones like allocation of resource, scheduling of data transfer, security as well as privacy and so on. Data transfer Scheduling is one of the significant issues for improving the proficiency of all cloud based administrations. In cloud computing, data transfer scheduling is utilized to allot the task to best reasonable asset for execution. There are various types of data transfer scheduling algorithms. A few issues like execution time, execution cost, high delay time, complexity, and high data transfer cost as well as various optimization problems have been measured in existing papers. To tackle all the above problems, in this paper, a new Adaptive approach are introduced which is a combination of Monarch Butterfly and Genetic (AMBA) Algorithm based data transfer scheduling is proposed. So here the concept is to develop an optimal algorithm for scheduling the data transfer in an efficient way which helps in reducing the data transfer time. The performance of proposed methodology analyzed in terms of evaluation metrics.**

***Keywords*—Information transfer scheduling; AMBA; throughput maximization; migration operation**

I. INTRODUCTION

Cloud computing platform could be defined as the usage of computing assets for example Software as well as Hardware, Which the clients get them in type of administration through a system (regularly web). It goes for spreading enormous scale segments and assets that are required away, learning, calculation and data for scientific inquire about after some time, cloud applications are slanting more to rely upon system in the regions of intuitiveness or information access and furthermore their requests for prerequisites are expanding step by step. Calculation is mentioned by some specific errands which are referred to as employments and dictated by calculation, organize limit and capacity. Cloud applications utilize numerous VMs in the preparing of those information volumes.

Along these lines, such applications deal with various employments where they are done by accessible assets so the best results, briefest reaction time, most brief time of finishing and utilization of assets can be gotten (1) (2). It is a web based processing that gives assets respect to a compensation for each utilization premise. Because of the upsides of high figuring force, low administrations cost, better execution, adaptability, openness just as accessibility it has turned into a utility. It is separated into application, stockpiling and availability sections. Each section fills for different needs and gives items to organizations and peoples in the world. Without establishment retrieve their personal or official documents at any PC, it enables purchasers and organizations to access applications using internet. Virtualization is a basic component of distributed computing. It is programming that isolates physical foundations to make different assets (3) (4). The primary preferred position of job scheduling computation is to accomplish a superior figuring and the best framework throughput. Planning oversees accessibility of CPU memory and great scheduling strategy gives most extreme usage of asset (5).

The ruler butterfly improvement (MBO) calculation has demonstrated to be a successful apparatus to tackle different sorts of optimization problems. In any case, in the fundamental MBO calculation, the search methodology effectively falls into local optima, causing premature convergence as well as poor presentation on numerous complex optimization issues. It can diminish an arranging task in addition to improve the computational productivity (6). Overcoming MBO limitations, a selfish algorithm was introduced in genetic migration and genetic modification, as well as a method linked to the work of genetic scientists, maintaining a balance between geographical diversity and integration. Local (7). Getting a new position depends on how long you pay for a solar compass or an attractive compass, however most depend on how long the compass takes. New generation immigrants are produced by high magnetic forces and follow the previous generation where they go and the best features of any generation will continue to be passed on to the next generation (8) (9). FF can simply achieve the global optimum and it has solves the issues quick and it effectively flexible to the applications (10).

In this work, we are utilizing another new scheduling concept to transfer data in efficient manner dependent on Adaptive Algorithm. Our proposed adaptive algorithm powerfully solves the data transfer scheduling struggles. The aim is to build up a scheduling algorithm to transfer data utilizing a adaptive approach Consolidating monarch butterfly

as well as genetic algorithm (AMBA) in cloud computing platform it find out best scheduled path to reduce data transfer time. The MB is a recently used algorithm and its basic nature is to solve global operational problems very quickly, and this algorithm is perfectly suited to similar processing and is well suited to the trade-off between durability and variability. Butterflies are in better health than their parents. This improves performance and speeds up the efficiency of data transfer planning. Lastly, data editing performance is analyzed based on different test metrics.

The subsequent content is in the order of; the proposed approach based literature survey is given in Section 2 and information transfer scheduling model is given in Section 3. The proposed optimal data transfer scheduling is explained in Section 4 and result and discussion is given in Section 5 and the conclusion part is in Section 6.

II. Literature Survey

Alex X. Liu (11) uses multiple immensity data transfers scheduling. In this bulk data migration among data centres was frequently a significant stage in deploying new services, improving dependability underneath failures, or executing an assortment of cost reduction methodologies for cloud organizations. These immensity amounts of data transferring consume enormous transfer speed as well as, further cause extreme system blockage. To beat these above downsides, here, they explored the Multiple Bulk Data Transfers Scheduling (MBDTS) issue to diminish the system blockage or network congestion. Transiently, they applied the store-and-forward exchange mode to lessen the pinnacle traffic load on the connection.

Roman Barták (12) Roman Barták (12) utilizes MAPF it manages the issue of finding a collision free path for a lot of operators. A Scheduling Based Approach to Multi Agent Path Finding with Weighted as well as Capacitated Arcs The real inspiration for the scheduling model of MAPF was its ability to normally incorporate different limitations. They considered especially the issues, where the limit of arc scan was more prominent than one that is more specialist output utilizes a similar curve in the meantime, and the lengths of circular segments was greater than one that is moving between various sets of nodes takes various occasions' times. These augmentations make the model nearer to reality than the original MAPF formulation.TevfikKosar (13) utilizes Data-aware scheduling in grid computing. This was Efficient as well as dependable access to enormous scale information sources along with documenting goals brings new challenges in widely scattered computing environment. The deficiency of the conventional frameworks as well as existing CPU-situated cluster schedulers in tending to these difficulties has yielded another emerging era: data aware schedulers. Here, they examine the limitations of the conventional CPU in handling the difficulty of demanding data management in wide ranging of distributed applications. Saurabh Kumar Garg (14) HPC clients need the capacity to increase quick and adaptable accessing to high performance computing abilities. Cloud computing guarantees to convey such kind of infrastructures utilizing data centres, by these HPC clients can use applications as well as access the information through Cloud from anywhere in the world. In any case, due to increase in the demand which drastically expands the vitality utilization of data centres, which has turned into a basic issue. High vitality utilization not just means high vitality cost which was decreased the overall revenue of Cloud suppliers, yet in addition high carbon emanations which was not suitable for environment. To tackle this problem, they introduced near-optimal scheduling strategies that exploit heterogeneity over various server farms for a Cloud provider.Yuan Zhang (15) uses a planning strategy that consists of two sections. With the allocation of a computer component, which is part of a long-term job measurement, planning can divide the work process into task categories by mandatory data transfer, because each Task Team is assigned a computer process. The unit that completed the task team at the most convenient time. Instead of simulation and wireless communication, they simply use the robin circular rule. The process of measuring the phase release was discussed, according to the performance phase analysis. In this case, the outcome indicates that the proposed resource planning strategy may be delayed.

III. Information Transfer Scheduling

Information transfer scheduling is used to transfer the information from source to destination leads to minimum time and maximum throughput. In the cloud, information is stored in different nodes, not all information is stored in local servers, and some of the nodes may have to fetch information from distance servers. In this case, time may be increased. So, to avoid time consuming, replicas are generated. In this manner, needed information can be fetched from one of the replica servers. In the information retrieval process, initially, the node is chosen for information recovery, an information transfer path is specified from the requesting node to the information transfer node. A lot of paths are available for one transmission. Among them, we will choose the one shortest path which wills lead to minimize the time and maximize the throughput. If we select the multi-path means, the system will suffer from high jitter. A naive strategy is to choose hubs and paths arbitrarily, however, it might outcome in overwhelming congestions on certain connections, prompting long information recovery time, since it not consider link bandwidths and the overlaps of chosen paths and hubs. Information transfer based on the randomly chosen path is given in Fig. 1. To overcome the problem the shortest part is optimally selected with the help of hybridization approach. The optimal information transfer scheme is given in Fig. 1.

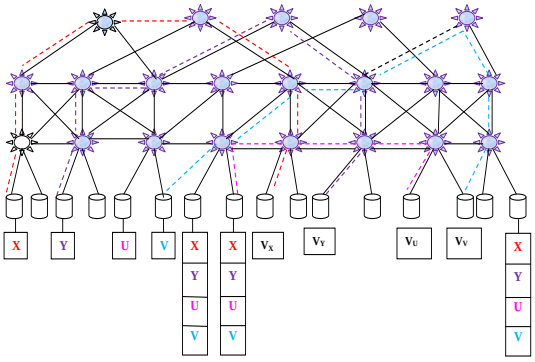


Fig. 1. Information Transfer based Randomly Chosen Path.

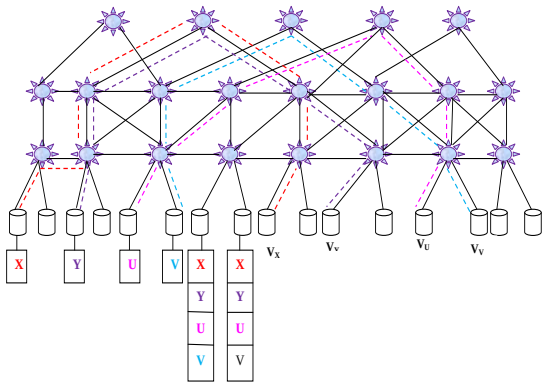


Fig. 2. Optimal Information Transfer Scheme.

In Fig. 2, X, Y, U and V are the information and V_x, V_y, V_u *and* V_v are retrieving hubs. Hub V_x retrieving information x which is specified as red dash line, V_y is retrieving information y which is specified as violet dash lines, V_u is retrieving a information u which is represented as pink dashed lines and V_v retrieving a information v which is specified as blue dashed lines. In Fig. 2, both data transfers share common traffic links at high volume and can lead to higher transmission times for the lower 2 numbers to pass in batches, resulting in shorter data retrieval. It is still distributed among computing nodes, or all data is accessible, so few sites may need to access data from remote locations. In this case, the requested information can be obtained at one of its locations. At that time, when a node is preferred for data acquisition, the request method for that requesting node requires specification of the data transfer. Let as consider an example of individual system, in a polygon as shown in Fig. 1, the four information elements (X, Y, U and V) are stored in 4 duplicates, and each link has a bandwidth of data per second. Note that it takes at least a second to move data between two centers.

Take simultaneously, the node will retrieve data V_x, retrieve data **x,** V_y retrieve data **y,** V_u, and retrieve data V_v, both active and random can have a information retrieval time of 4 seconds; although a good solution takes a second. The idle method works poorly, because all data transfers go through the normal connection, causing the bottle to be shown in Fig. 1. Less time was taken by correct solution because node selection and all data transfers with sets Links are not shown in Fig. 2 Network (DCN) shown in Fig. 3.

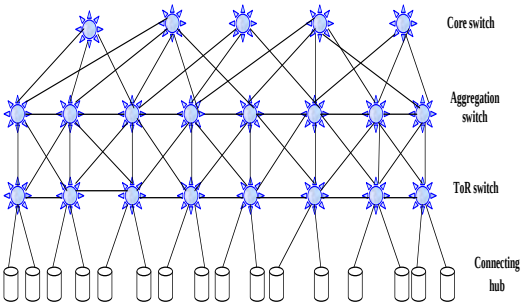


Fig. 3. Information Centre Network Topology.

IV. Proposed Information Transfer Scheduling System

In this work, we have intended to enlarge an information transfer scheduling scheme based on Adaptive Algorithm. This adaptive approach finds the best path based on the least information transfer time-the maximum throughput. The aim is to develop an optimal information transfer scheduling using an Adaptive approach combining monarch butterfly (MB) and genetic operators (AMBA) in the cloud computing environment. The proposed methodology structure is given in Fig. 4.

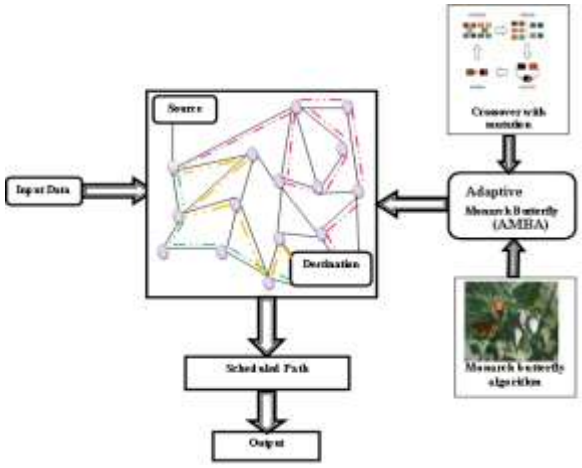


Fig. 4. Proposed Methodology Structure.

The proposed MBA is a recently developed optimization algorithm which is mainly used for solving global optimization problem (1). Basically, the MBA has two operators, namely migration and adjusting. But in basic MBA, we have some difficulties in the search process conducted by the butterfly adjusting operator. Early stages of algorithm's execution, the search process that is exceedingly directed towards the current best solution in the population, in some of the algorithm's run generate poor results. To tackle with this deficiency, MBA is adaptive with genetic operators. The step by step process of AMBA based optimal path selection process is explained below:

Step 1: Initialization phase

Solution initialization is a chief process for the entire optimization problem. Here, the initial solution is initialized at random. Initially, we initialize the solution parameters. For data retrieval process, N number of the path is available. To increase the speed and minimize the transformation time and cost, we optimally choose the path. In this paper, the path is considered as a monarch butterfly (solution) and the population is considered as a butterfly. Initially, monarch butterflies are randomly initialized.

Consider information x and y is transferred to hub u and v. Here, for security purpose, we have generated three replicas for both information x and y. The system generates multiple. The initial solution format is given below:

$$S_i = \left(P_1, P_2, \ldots, P_N\right) \quad (1)$$

After that, we divide the population into two groups namely subpopulation 1(SP_1) and subpopulation 2 (SP_2). NP represented the total number of monarch butterflies. The number of MB's present in SP_1 is calculated using equation 2.

$$NP1 = ceil(p \times NP) \quad (2)$$

The number of MB's present in SP_2 is calculated using equation 3.

$$NP2 = NP - NP1 \quad (3)$$

Where;

p -> migrating speed of monarch butterflies with $p = 5/12$ in MB,

Ceil(x) rounds x to the nearby integer larger than or equivalent to x,

Step 2: Fitness calculation

After the solution initialization, the fitness of each butterfly is calculated. The fitness function is based on cost and time. If the butterfly attains the minimum cost and time means that the butterfly is considered as the best fitness.

$$Fitness = \min(\text{cost}, time) \quad (4)$$

Step 3: Migration operation

After the fitness calculation, each butterfly migrates their position. The migration function can be written as follows:

$$S_i^{q+1} = \begin{cases} S_{r1}^q, r \le p \\ S_{r2}^q, r > p, \end{cases} \quad (5)$$

$\boldsymbol{S_i^{q+1}}$ is a k-th part of $\boldsymbol{S_i}$ in generation $\boldsymbol{q+1}$; similarly, $\boldsymbol{S_{r1}^q}$ denotes the k-th fraction of $\boldsymbol{S_{r1}}$ in generation **q,**, and $\boldsymbol{S_{r2,k}^q}$ q is the k-th fraction of $\boldsymbol{S_{r2}}$ in generation q; the current generation number is q, and the monarch butterflies r1 and r2 are randomly selected from subpopulation 1 and subpopulation 2, so here, r is calculated by $\boldsymbol{r = r \times peri,}$ where peri is the time of migration, which is 1.2 in MB and the rand is a random number in (0, 1).

Step 4: Adjustment operation

After migration operation, adjustment operation is done in SP_2.The subsequent formula is described as,

$$S_i^{q+1} = \begin{cases} S_{best}^q rand \le p \\ S_{r3,}^q rand > p, rand \le BAR \\ S_i^{q+1} + \alpha \times (ds_k - 0.5), rand > p, rand > BAR, \end{cases} \quad (6)$$

In any case, S_i^{q+1} is part of the k-th of S_j in generation q + 1; Similarly, S_{best}^q is the k-th part of S_{best} in generation q, which is the best place for monarch butterflies in world 1 and land 2, $S_{r3,k}^q$ part S_{r3} in generation q, king butterfly r3 selected at random in the case of less than 2, and BAR is the conversion or correction rate, if BAR is less than the random number r and the k-th fraction of x_j at $q+1$, when σ is measured, and $\sigma = S_{max} / q^2$, where S_{max} is the highest travel step.

Step 5: Crossover operator

After the migration process, to improve the MBA, an additional operator is integrated with the MBA. Crossover is the process by which genes are selected from the chromosomes of parents and new offspring. Crossover can be done with binary code codes, coding code, pricing code and encoding Fig. 5.

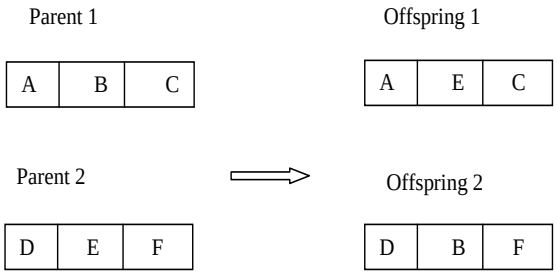


Fig. 5. Crossover Process.

Step 6: Mutation operator

After skipping a task, the solution is updated with the help of modifications. Genetic modification function, can search for new locations in contrast to the crossing. The crossover is called the exploit operator and the conversion is a form of proof Fig. 6.

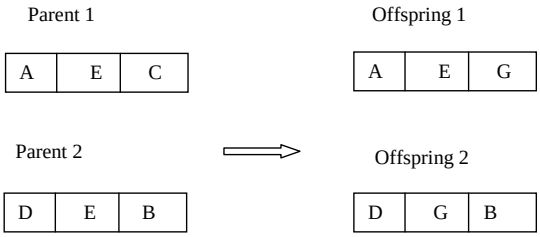


Fig. 6. Mutation Operator.

Step 7: Termination criteria

The algorithm stops its use only when selecting the highest frequency selection and the solution contains the best body weight and is given as the best option. The AMBA based optimal path selection algorithm is given below.

Algorithm

Input: parameter of MBA, Cross over rate, mutation rate, number of hub, sender, destination, number of replicas

Output: Optimal path

Start

1. Randomly initialize the number of paths available in data transfer
2. Divide the whole population into two SP1 and SP2
3. Calculate the fitness for each path (monarch butterfly)
4. While k<MaxGen do
5. Sort all individual in the population based on the fitness value
6. For i=1 to SP1 (all butterflies in the SP1) do
7. Apply butterfly migration operation to get a new butterfly
8. End for
9. For j=1 to SP2 (all butterfly present in the SP2) do
10. If t<Maximum generation 0.5 then
11. Generate new butterfly in SP2 by using cross over and mutation operator
12. Else
13. Generate new butterfly in SP2BY using butterfly adjustment operator
14. End if
15. End for
16. Merge SP1 and SP2
17. Increase the iteration counter k by one
18. End while
19. Return bet butterfly in the whole population.
20. Output

Optimal path

V. Simulation and Results

In this section, the proposed information transfer scheduling scheme based on adaptive Algorithm. Our proposed adaptive algorithm efficiently solves information transfer scheduling problems. In this paper, we used an adaptive approach which integrating the monarch butterfly algorithm and Genetic algorithm (AMBA) in the cloud computing environment.

A. Experimental Results

While testing the AMBA algorithm performance, it may be difficult to verify the functionality of all the same algorithms. The main use of this to solve the problems of speeding up the earth again, this algorithm is perfectly fit for similar processing and can trade between durability and variability. In the proposed work, finally, the effectiveness of the data transfer planning process is analyzed according to different test criteria. Comparing to the existing optimization techniques our proposed AMBA achieves the better results. The following Fig. 7, 8, 9 and 10 shows the waiting time, turnaround time, response time and fitness of the proposed approaches.

Analyzing Fig. 7, 8, 9, and 10 show the comparative analysis of proposed against existing based on waiting time, turnaround time, response time and fitness. Analyzing Fig. 7 our proposed AMBA algorithm achieves the minimum waiting time of 865, 1245, 1625, and 2248 for other existing optimization and without optimization algorithms because Initially it calculates the jobs entirety completion time on every computing node, as well as file access time or duplication time among stored files along with every Computing node file size, bandwidth and the waiting time for every job earlier than it process and processing time. The above Fig. 7 clearly specifies our proposed approach waiting time is minimum for comparing all the other existing systems. Fig. 8 turnaround times is the total time taken among the whole process which is measured by the time interval from starting time and completion time of the process. Analyzing Fig. 8 our proposed AMBA algorithm achieves the minimum time of 1985ms, 2635ms, 3845ms and 4958ms for other existing MBO, GA and without optimization techniques.

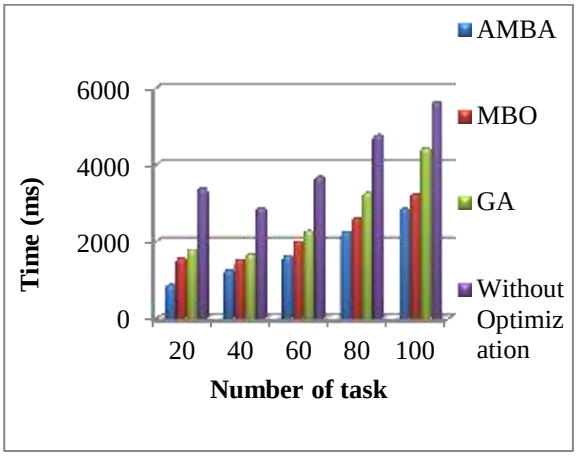


Fig. 7. Waiting Time Analysis.

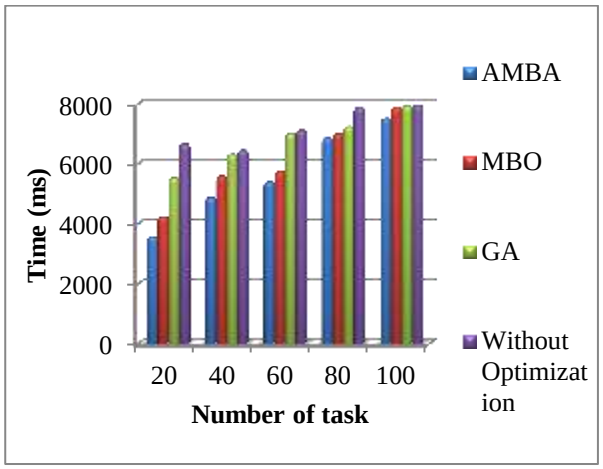


Fig. 8. Turnaround Time Analysis.

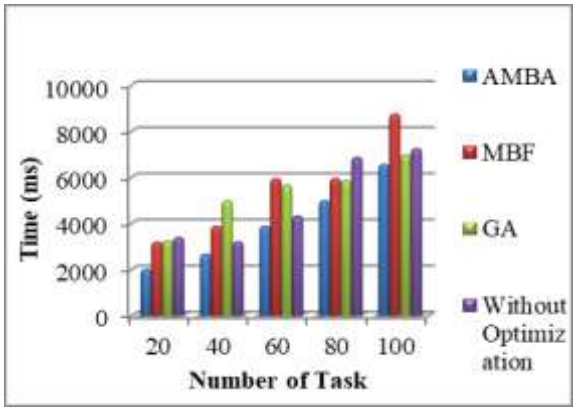


Fig. 9. Response Time Analysis.

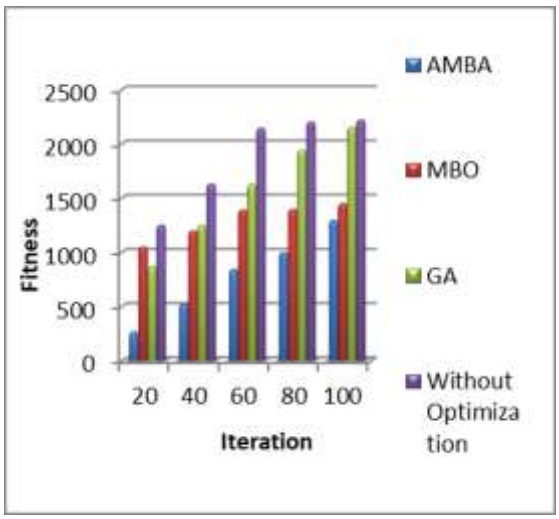


Fig. 10. Fitness versus Iteration.

Fig. 9 shows the comparative analysis based on AMBA it minimizes the response time. Our proposed AMBA achieves the minimum response time of 3545ms, 4854ms, 5365ms, and 6847ms for other existing optimization and without optimization algorithms because the time taken to process a user request from the source point it is made up till the destination point it is received. Comparing to all the other existing techniques, the above Fig. 9 clearly specifies our proposed approach response time is minimum for comparing all the other existing systems.

Fig. 10 our proposed AMBA minimize the objective function value of 1985, 2635, 3845, and 4958 for other existing optimization and without optimization algorithms.

Comparing to all the other existing techniques, AMBA it minimizes the fitness of objective function values. Experimental studies have shown that our proposed AMBA is better than other available solutions for full durability and delay because genetic operators are used in MB in migration operation, and this built-in strategy resides only with people with better potential than their parents. The performance will be improved and it speeds up the efficient data transfer process. In the proposed project, a genetic operator is used in MB to do migration work, and this built-in strategy can only accommodate monarch butterfly people who are more highly trained than their parents. Computer results indicate that the proposed process exceeds existing methods. Individual power is used as the purpose function of the solution for the same system solution. At the same time the solution reduces the amount of meaningful work, the best that can be achieved. From the test results, we are well aware that our proposed method achieves better results compared to existing methods.

VI. Conclusion

In this paper, a new adaptive Monarch Butterfly and Genetic operator (AMBA) Algorithm could be proposed, make use of this algorithm the information transfer scheduling problems could be solved efficiently. The aim of research is to improve the best data transfer planning using AMBA on a cloud platform where you will find the most optimized method which leads to the shortest time of data transfer; in other words, the highest number. In this case, MB is one of the proposed algorithms that will be used later. The main application of this to solve the problems of re-accelerating the earth, this algorithm is perfectly suitable for the same processing and is able to trade between durability and variability. The genetic operator can easily access global operations and solve problems quickly and easily adapt to applications. In the proposed work, the GA applies methyl bromide in migration operations, and this included strategy can only accept monarchs that are healthier than their parents. The performance will be improved and it speeds up the efficiency of data transfer planning. Finally, our proposed approach outperformed other existing MBO, GA and without optimization approaches based on waiting time, TAT, response time as well as Fitness.

References

[1] EsmaYildrim,EnginArslan,Jangyoung Kim, and TevfikKosar,In IEEE," Application level optimization of big data transfer" IEEE, March (2015).

[2] Aakanksha Sharma, Sanjay Tyagi, "Task Scheduling in Cloud Computing" International Journal of Scientific & Engineering Research, Volume 7, Issue 12, December-2016.

[3] D. A. Agarwal and S. Jain, "Efficient Optimal Algorithm of Task Scheduling in Cloud Computing Environment," International Journal of Computer Trends and technology (IJCTT), 2014.

[4] M. Kalra and S. Singh, "A review of Metaheuristic Scheduling Techniques in cloud computing," Egyptian Informatics Journal, Elsevier, 2015.

[5] E. Kumari and M., "A Review on Task Scheduling Algorithms in Cloud Computing," International Journal of Science, Environment and Technology, 2015.

[6] LinSun, JiuchengXu, "Improved Monarch Butterfly Optimization Algorithm Based on Opposition-Based Learning and Random Local Perturbation", 10 February 2019.

[7] Shifeng Chen, Rong Chen, "A Monarch Butterfly Optimization for the Dynamic Vehicle Routing Problem", Department of Information Science and Technology, Dalian Maritime University, 12 September 2017.

[8] Kaushik Kumar Bhattacharjee, S.P. Sarmah, "Monarch Migration Algorithm for Optimization Problems", Department of Industrial & Systems Engineering, Proceedings of the 2015 IEEE IEEM.

[9] P. Guerra, C. Merlin, R. Gegear, and S. Reppert, "Discordant timing between antennae disrupts sun compass orientation in migratory monarch butterflies," Nature Communication, vol. 3, p. 958, 2012.

[10] Umar SHEHU, Safdar, Gregory, "Fruit Fly Optimization Algorithm for Network-Aware Web Service Composition in the Cloud", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[11] Sen Su, Zhongbao Zhang, "Multiple bulk data transfers scheduling among data centres",Department of Computer Science and Engineering, Michigan State University,18 February 2014.

[12] Roman Barták, Marec VLK, "A Scheduling-Based Approach to Multi-Agent Path Finding with Weighted and Capacitated Arcs", Scheduling and Planning, AAMAS July 10-15 2018.

[13] TevfikKosar, Mehmet Balman, "A new paradigm: Data-aware scheduling in grid computing", Department of Computer Science, Elsevier 2009.

[14] Saurabh Kumar Garg, Chee Shin Yeo, "Environment-conscious scheduling of HPC applications on distributed Cloud-oriented data centers", Cloud Computing and Distributed Systems Laboratory, Department of Computer Science and Software Engineering, Elsevier 2010.

[15] Yuan Zhang, Nanjing, "Resource Scheduling and Delay Analysis for Workflow in Wireless Small Cloud", IEEE Transaction on Mobile Computing, volume: 17, Issue: 3, March 1, 2018.

Robotic Education in 21st Century: Teacher Acceptance of Lego Mindstorms as Powerful Educational Tools

Mardhiah Masril[1]*, Ambiyar[2], Nizwardi Jalinus[3], Ridwan[4], Billy Hendrik[5]
Dept. Faculty of Computer Science, Universitas Putra Indonesia "YPTK" Padang, Padang, Indonesia[1, 5]
Dept. Faculty of Engineering, Universitas Negeri Padang, Padang, Indonesia[1, 2, 3, 4]
Dept. Institute IR 4.0, Universiti Kebangsaan Malaysia, Bangi, Malaysia[5]

***Abstract*—Acceptance of robotic technology in education is a crucial issue in the revolution industry 4.0 era. This study aims to explore the acceptance of Lego Mindstorms Ev3 as one kind of robotic technology by the teachers as a learning resources that can develop teachers and student's skills. The Technology Acceptance Model (TAM) was introduced by using questionnaires. The questionnaires were responded by 22 elementary school teachers who have experiences with Lego Mindstorms ev3 kits in a workshop. The data was carried out by presenting descriptive statistics, correlation, and regression analyses. Based on the acceptance testing of Lego Mindstorms Ev3 with the TAM model, the result showed that subjective norms (SN) and self-efficacy (SE) as external variables were effective on the acceptance of Lego Mindstorms Ev3 as a learning tools by teachers. Teacher's SN have a positive correlation with perceived usefulness (PU), perceived ease to use (PE), and behavioral intention to use (BI). Teacher's self-efficacy were significant in predicting PE and BI. PU and PE had a positive effect on Attitude toward using Lego Mindstorms Ev3 by teachers, and it continued to use. Finally, most teachers have shown positive reactions to Lego Mindstorms Ev3 as educational tools.**

***Keywords*—Education; TAM; teacher acceptance; Lego; Robotic**

I. Introduction

The Acceptance of technology in the educational sector as innovative learning tools has become a favorite topic for exploration in a recent years. Learning tools have an important role in achieving the learning objectives especially technology-based learning tools [1], even technology-based teaching and learning facilities have an important role in transforming education [2]. One of the innovative learning tools in the industrial revolution 4.0 era is robotic technology. Most individuals seem to agree that robotics as a learning tool has provided many benefits in improving cognitive abilities, creative thinking skills [3], [4], problem-solving skills, collaboration skills, STEM, and computational thinking of students [5], [6]. There are many benefits of applying robotic technology in schools; therefore, introducing this technology early to students is important. The current students are Z-generation and alpha-generation, they are very easy to accept and adapt to the robotic technology that has been applied in their learning processes in the classroom. Student's acceptance of robotic technology in education has been demonstrated in previous research, students preferred to ask a robot about the information they wanted to know rather than ask an adult [7], students were more likely to followed behavioral suggestions offered by an autonomous social robot [8], the students familiar and had positive attitudes towards social robots (Fanuc LR Mate 200 ID, Sputnik, Nao) [9]. In addition to social robots, one of the educational robots that have high interaction with students is Lego. Lego is a kind of robotic technology widely used in the teaching-learning process [10]. In previous studies about student's acceptance of Lego Mindstorms, the young students (11-18 years) were more receptive to Lego Mindstorms in the learning process than the old students (19-24 years) [11], the early adolescents perceived educational and learning of robotics (Lego Mindstorms) as a source of employment, and as a way to high technology [12]. But teacher's acceptance of robotic technology also needs to be explored.

The teacher's acceptance of robotic technology as learning tools was very important to analyze because (1) The effective use of technology in classrooms is based on the attitudes of teachers to technology. Previous studies have shown that the attitudes of teachers, as well as expertise and skills in the use of technology, major factors influencing their initial adoption of technology and their future computer use actions [13]. The effectiveness of a teaching method is closely related to how teachers able to use technology to engage the learners [14]. Teachers' acceptance of technology is an important factor that influences the teacher's teaching method, teachers' behavior in the classroom and influences students' learning as well [15]; (2) robotic technology as innovative learning tools would not be effective if the teachers were not able to use it properly, a generation gap between teachers and students resulted in the teachers were more difficult to adapt to new technologies than students, the main characteristic differences among X, Y, and Z- generations are the mastery of information and technology; (3) Besides, teachers are responsible for ensuring that the technologies work correctly [16], demonstrating their added value in the teaching process, and offering a wider view of the purpose and significance of using technology.

A. Technology Acceptance Model

The Technology Acceptance Model (TAM) has become the most generally accepted theoretical paradigm for the acceptance of research technology. Introduced by Davis

*Corresponding Author

(1989), TAM was an evolution of the "Theory of Reasoned Action" [13]. TAM is a blueprint for how it will be adopted and used by technology users [17]. Another opinion said that TAM gave a recommendation when people use new technology for their activity based on usefulness and ease to use, this recommendation will influence their decision, accepting or rejecting it [18]. Original TAM by Davis [19] is shown in Fig. 1.

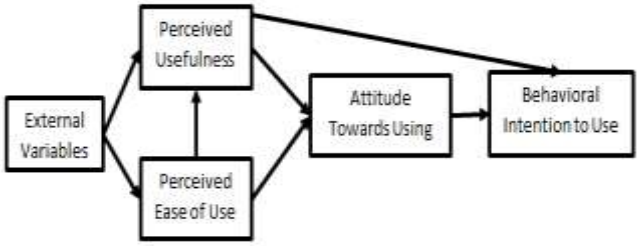


Fig. 1. Technology Acceptance Model (TAM).

Original TAM consists of two key variables (1) perceived ease of use (PE); (2) perceived usefulness (PU) [20], while attitude towards using, behavioral intention to use, and actual usage of the device are the deciding variables of technology acceptance [14], [21], [22], [23].

PU indicates the degree to which a person believes his/her success will be improved by the use of a system [14], [17]. PE indicates the degree of ease of use of the technology if the technology is too difficult to use, instead of using new technology, users will find the alternative form [17], [21]. PU and PE decide users' attitudes towards technology [24], [25], PU, and AT have a direct influence on behavioral purpose to use (BI) [19]. Intention to use is determined by attitudes and usefulness, and intention to use has a relationship with actual use [26], it can be determines, technology acceptance [27]. Several researchers have replicated Davis's original model [28] and added external variables to the model.

In this analysis, TAM was selected as the research model because it can assess the effect of external variables on perceptions, intention to use, and TAM as the best model for predicting user behavior towards new technology [29]. It was in accordance with this study which aims to determine the external factors that affect the behavior intention of teachers in using robotic technology in their class.

B. Acceptance of Robotic Technology by Teachers

Previous studies have been conducted on teacher acceptance for robotic technology, the potential benefits of the movements of the robot, and the significant correlation between changes in familiarity and perception in human-robot interaction have been identified [30]. The Thymio Robot application has high usability, teachers are very interested and have a desire to learn and master the use of the Thymio robot, through understanding new technologies, develop teacher professionals [10]. Teachers' ability to use robotic technology in training will enhance skills, self-confidence, and interactions between teacher and student, leading to a willingness to implement educational robotics in schools, this research shows the value of high-quality professional development in the self-efficacy of educators with the use of Educational Robotics, and suggests that new tablet-based wireless robotics platforms, such as LEGO® WeDo 2.0, enable younger learners to engage with this technology [31]. Another study reported that usage of WeDo as a robotics kit in the teaching-process helped the teachers build their confidence and knowledge to introduce students to computational thinking [32]. Reich-Stiebert explored the ability of teachers to use robots in different learning settings in a German survey of 59 teachers. Their findings revealed teachers' very negative attitude toward educational robots. The authors concentrated on the robot NAO as an assistant to teachers in this report. The research was distinguished by age, gender and subject taught. There was no major effect of age and gender on attitudes. The topic taught, however, had a substantial impact: teachers chose to use robots in STEM-related domains [33].

C. Objectives of the Present Study

Few studies have reported on Lego / robotic technology acceptance in the learning process, but the majority of studies focused on students' interaction with robotic technology in the classroom. Teachers' acceptance more discussed robots as an assistant of a teacher in the classroom, and very little information available on teachers' acceptance of Lego Mindstorms Ev3 as learning practices in the classroom while Lego practices were widely used in the teaching-learning process.

The purpose of this study was to explore the teachers' acceptance of robotic technology, especially Lego Mindstorms Ev3 as learning tools by elementary school teachers. To evaluate the teacher's acceptance, we used the Technology Acceptance Model (TAM) with external variables were subjective norm and self-efficacy (adopted by Yuen, 2008).

The behavioral decisions or intentions of individuals are always influenced by the other people around them [34]. The subjective Norm, a person's subjective norm is his or her perception that most people who are important to him or her think he or she should or should not perform the behavior in question [13]. Theory of Reasoned Action claimed that motivation to comply (known as compliance) is a predictor for subjective norms, and subjective norms are a predictor for intentions [35]. Subjective norm is one of the main variables in the Theory of Planned Behavior (TPB) [36] that can influence behavioral intentions in IT adoption.

Self-efficacy and technology have a strong relationship. Bandura defines self-efficacy as "People's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances [37]. High-self efficacy can aid individuals in initiating cross-cultural interactions, persisting in the face of early failures, and engaging in problem-solving as a way of mastering necessary skills [38]. Self-efficacy is one of the key drivers of human activity and it has been found to have both direct and indirect impact on the intention and actual use of different technologies [39]. Self-efficacy can have effects on individual intentions [40], individual engagement, and behavior.

The result of this study could be useful to help teachers adopting Lego Mindstorms Ev3 for their teaching practices, may provide information on how teachers accept robotics

technology as a learning tool in schools, and improve educational sectors to adapt to technological developments.

D. Hypothesis

Based on Fig. 1, the following hypotheses of this study:

H1: The perceived ease of use (PE) has a positive effect on the perceived usefulness (PU) of Lego Mindstorms Ev3 as educational tools by teachers.

H2: The perceived ease of use (PE) and the perceived usefulness (PU) have a positive effect on the Attitude toward using (AT) of Lego Mindstorms Ev3 as educational tools by teachers.

H3: The perceived usefulness (PU) and attitude toward using (AT) have a positive effect on the behavioral intention to use (BI) of Lego Mindstorms Ev3 as educational tools by teachers.

H4: The Subjective norm (SN) has a positive effect on the perceived usefulness (PU) of Lego Mindstorms Ev3 as educational tools by teachers.

H5: The Subjective norm (SN) has a positive effect on the perceived ease of use (PE) of Lego Mindstorms Ev3 as educational tools by teachers.

H6: The Subjective norm (SN) has a positive effect on the behavioral intention to use (BI) of Lego Mindstorms Ev3 as educational tools by teachers.

H7: The Self-efficacy (SE) has a positive effect on the perceived usefulness (PU) of Lego Mindstorms Ev3 as educational tools by teachers.

H8: Self-efficacy (SE) has a positive effect on the perceived ease of use (PE) of Lego Mindstorms Ev3 as educational tools by teachers.

H9: Self-efficacy (SE) has a positive effect on the behavioral intention to use (BI) of Lego Mindstorms Ev3 as educational tools by teachers.

II. Methodology

A. Procedure

1) Teachers actively interacted with Lego Mindstorm Ev3 in a workshop (attending 2 days of workshop).

2) After the workshop was completed, the teachers were asked to fill out a questionnaire related to teachers' acceptance of the use of Lego Mindstorm Ev3 as learning tools.

3) To evaluate the teacher's acceptance, we used the Technology Acceptance Model (TAM) developed by Davis (1989).

4) The reliability test used Cronbach's alpha for each item, the alpha value is at least 0.7 and higher, it was mean reliable [41].

5) Analyzing the data descriptive statistic to determine the minimum, maximum, average, and standard deviation values.

6) Hypothesis testing used regression analysis.

7) Data analysis was performed with SPSS software.

B. Participant

Participants in this study consisted of 22 teachers from five elementary schools in Padang, West Sumatra, Indonesia, who had attended a workshop on robotic technology. The teacher's characteristics can be seen in Table I.

C. Workshop of Robotic for Teacher

This workshop has been held for elementary school teachers, so that teachers could interact directly with Lego Mindstorms Ev3 as learning tools in this workshop so that teachers can ensure ease of use and benefit for teachers and students in the learning process. Lego Mindstorms offers an environment for teachers and students to interact in an exciting, creative way [42]. The workshop was held for 2 days; detailed activities are shown in Table II.

The first section, the introduction of robotic technology, the implementation of robotic technology especially in the education sector, and the benefits of applying robotic technology for students. Information presented to teachers via an interesting audio visual media, these activities focused on gave deeper knowledge of robotic technology.

The clarification preceded by getting to know the Lego Mindstorms Ev3 as the methods to be used in this workshop as an experiment. The components of the Lego Mindstorms Ev3 package and the tasks of each of these components are introduced to the teacher. Lego Mindstorms Ev3 consist of building kits and a programmable control unit that can allow a robot to be built. This kit includes all essential components, such as connectors, axles, bushings, beams, frames, tubes, gears, belts, shafts, wheels, motors, sensors, and control centers, necessary for the construction of a robot [4]. After this section, the teachers were able to understand the function of Lego Mindstorm Ev3 kits.

TABLE I. Demography Table of Participant

Characteristic	Value	Frequency
Age	< 30 30-35 > 35	3 11 8
Gender	Men Women	10 12

TABLE II. Activities of Robotic Technology Workshop

Sections	Activities	Duration
Section 1	Introduction of robotic technology and Lego Mindstorms Ev3 kits.	(at 08.00 am to 11.50 am on the first days)
Section 2	Project 1 – Create Tracker Tank Bot	(at 01.00 pm to 05.00 pm on the first days)
Section 3	Project 2 - Create humanoid robot Ev3rstorm	(at 08.00 am to 11.50 am on the second days)
Section 4	Project 3 - Create the robots based on teacher's creations	(at 01.00 pm to 05.00 pm on the second days)

In the second section, via the Lego Mindstorms Ev3 kits, the teachers were able to get new experiences. The search for teachers to create groups started with this segment (a group consisting of two teachers). A Tracker Tank Bot was developed for the project-1, this project started with how to design, develop, build, before how to control a Tracker Tank Bot. In the third segment, each group of teacher's plan, create, build, and learn how to program Brick as a control center in Lego Mindstorms Ev3 in project-2. A programme for controlling sensors (color, ultrasonic, contact, infrared, gyro, temperature sensor) [43], and motors can be sent by the brick as the robot actuator.

The teachers were asked to make robots based on their creations in the last sections of project-3, this activity offered the teachers an opportunity to develop their ideas to build a new robot or change a robot through Lego Mindstorms Ev3.

D. Instrument

The questionnaires were used to obtain information for this quantitative study. The instrument of questionnaire adopted by prior studies and modified to be compatible with this study context. The questionnaire was consist of 19 items, four items for perceived usefulness (PU), four items for perceived ease to used, three items for attitude (AT), four items for intention to use (BI), were adopted from Davis [44], and Çukurbaşı [45]. Two items for subjective Norm (SN), and two items for self-efficacy (SE), were adopted from Yuen [13], Bröhl [28], and Nadlifatin [36]. All items were measured in a 5-point Likert Scale with 1 as "strongly disagree" and 5 as "strongly agree". Details of all items used in the questionnaire are provided in Appendix A.

III. RESULT

This section has presented the results of descriptive statistics, correlation analysis, and regression analysis. Cronbach's alpha was calculated to test reliability (an alpha of at least 0.7).

According to Table III, all items of questionnaire were reliable, these were indicated by alpha coefficient 0.890. Correlation between variables were presented in Table IV and Fig. 2.

Table IV presented high correlation between SN and BI (0.843), medium correlation between: PE and AT (0.620); PU and AT (0.573); AT and BI (0.580); SN and PU (0.438); SN and PE (0.561); SE and PE (0.544); SE and BI (0.527), no correlation between PE and PU; PU and BI; SE and PU.

According to Table V, a p-value of PE to PU = 0.064 > 0.05, indicated that PE has not had a positive effect on PU of Lego Mindstorms Ev3 as learning tools. This means that the ease of using Lego Mindstorms Ev3 as learning tools has no effect on the perceived benefits of teachers in the learning-process (H1 was rejected).

PE and PU simultaneously to AT showed a p-value of 0.001> 0.05, B-value of PE = 0.372 and PU = 0.341, it indicated the perceived ease of use (PE) and the perceived usefulness (PU) have positive effect on the Attitude toward using (AT) of Lego Mindstorms Ev3 as educational tools by teachers. Rated R = 0.714 with a coefficient of determination (R-Square) = 0.510, it could be stated that the PE and PU influence the AT of 51.0%, it can be stated that if Lego Mindstorms Ev3 as learning tools easier to use and the greater the benefits in the teaching and learning process in the classroom, the teachers are more able to accept LEGO Mindstorms Ev3 as learning tools (H2 was accepted).

B-value of AT = 0.965 and PU= -0.091, partially AT has a positive effect on BI while PU has not had a positive effect on BI, but PU and AT simultaneously to BI presented p-value of 0.019 < 0.05, it indicated the perceived usefulness (PU) and attitude toward using (AT) have positive effect on the behavioral intention to use (BI) of Lego Mindstorms Ev3 as educational tools by teachers. Rated R = 0.583 with a coefficient of determination (R-Square) = 0.340, it could be stated that the PU and AT influence the BI of 34.0% (H3 was accepted).

TABLE III. DESCRIPTIVE STATISTIC OF THE ITEM OF QUESTIONNAIRE

Items of questionnaire	Alpha (α)	Min	Max	Mean	SD
PU1	.894	3	5	4.32	.568
PU2	.889	3	5	4.50	.598
PU3	.891	4	5	4.68	.477
PU4	.884	3	5	4.73	.550
PE1	.876	3	5	4.32	.646
PE2	.887	3	5	4.09	.526
PE3	.890	4	5	4.59	.503
PE4	.884	4	5	4.86	.351
AT1	.877	3	5	4.18	.733
AT2	.886	4	5	4.55	.510
AT3	.887	4	5	4.64	.492
BI1	.875	3	5	4.14	.774
BI2	.881	3	5	4.23	.685
BI4	.890	3	5	4.27	.550
BI5	.886	4	5	4.59	.503
SN1	.877	3	5	4.23	.685
SN2	.878	3	5	4.27	.631
SE1	.880	4	5	4.45	.510
SE2	.884	3	5	4.45	.671

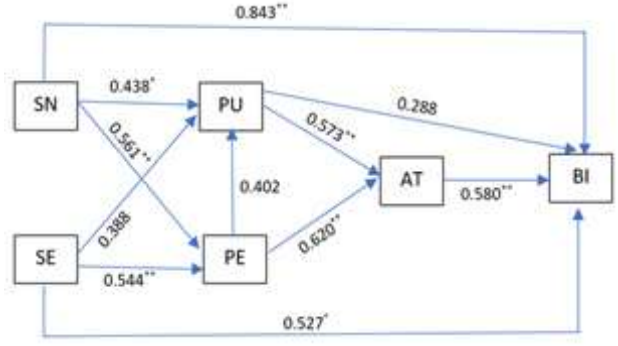


Fig. 2. Correlation between Variables of the Result.

TABLE IV. CORRELATION BETWEEN VARIABLES

Independent variables	Dependent Variables	Person correlation	*p-value*
PE	PU	0.402	0.064
PE PU	AT	0.620** 0.573**	0.002 0.005
PU AT	BI	0.288 0.580**	0.193 0.005
SN	PU	0.438*	0.042
SN	PE	0.561**	0.007
SN	BI	0.843**	0.000
SE	PU	0.388	0.074
SE	PE	0.544**	0.009
SE	BI	0.527*	0.012

TABLE V. HYPOTHESES AND REGRESSION SCORES

The model hypothesis	Independent variables	Dependent variables	B	t	p-value	R	R^2
H1	PE	PU	0.364	1.962	0.064	0.402	0.161
H2	PE PU	AT	0.372 0.341	2.647 2.204	0.001	0.714	0.510
H3	PU AT	BI	-0.091 0.965	-0.289 2.716	0.019	0.583	0.340
H4	SN	PU	0.527	2.179	0.042	0.438	0.192
H5	SN	PE	0.745	3.029	0.007	0.561	0.314
H6	SN	BI	1.400	7.007	0.000	0.843	0.710
H7	SE	PU	0.525	1.885	0.074	0.388	0.151
H8	SE	PE	0.813	2.903	0.009	0.544	0.296
H9	SE	BI	0.983	2.776	0.012	0.527	0.278

Subjective norm (SN) variable were found significant in predicting: PU (p-value = 0.042< 0.05); PE (p-value = 0.007 < 0.05); and also predicting BI (p-value = 0.000 < 0.05), it means H4, H5 and H6 were accepted. Self-efficacy (SE) variable was found not significant in predicting PU (p-value = 0.074 > 0.05) means H7 rejected, but SE were significant in predicting: PE (p-value = 0.009 < 0.05); BI (p-value = 0.012< 0.05) means that H8 and H9 were accepted. Table V indicated that out of nine hypotheses two of them were not accepted.

IV. DISCUSSION

The outcome of this study showed that teachers commonly agree that Lego Mindstorms Ev3 as a learning method was expressed in all questionnaire items averaging greater than 3.0. The perceived ease of use of Lego Mindstorms Ev3 as learning tools has not a positive impact on the perceived usefulness in this study, this is different from what has been written in the literature [13]. This study showed that elementary school teachers gave a positive attitude toward usage (AT) Lego Mindstorm Ev3 because there were relationship between PU and PE; (1) Usage of Lego Mindstorms Ev3 provided benefits to improve productivity, performance, efficiency teachers in classes, and the students were actively involved in the lesson; (2) Lego Mindstorms Ev3 as learning tools were quickly to understand and were easy to operate, this outcome was consistent with previous studies, that educational robotics improved teacher attitudes because robotics improved STEM interaction and teaching [46], The teachers indicated that the use of the Lego WeDo 2.0 robotics kit provides a unique opportunity for computer skills to be developed; it focuses on activities that facilitate problem-solving and group work with primary school students [32]. The relationship between perceived useful and easy to use for technology acceptance has been demonstrated for numerous information technologies [24].

The previous studies explained that attitudes toward teaching assistance robots mainly determined teacher's intended use for the robots [20], the highest positive effect was determined of attitude toward to intention of use of a Telepresence Robot in the classroom by teachers compared with other variables (perceived usefulness, perceived enjoyment, trust of technology, social influence, and gender) [47], it was in accordance with this study indicated that Lego Mindstorms Ev3 as learning tools were attractive, fun, and useful in the learning process had a positive influence on the behavioral intention of the teacher to use Lego Mindstorms in a lesson.

Furthermore, by emphasizing one of the social factors such as subjective norm, this study showed that organizations and other teachers supported to use of Lego Mindstorms Ev3 in the classroom, subjective norms have a positive correlation with PU, PE, and BI, this result consists to a previous study that one of important predictors in the robot acceptance model was subjective norm [48].

The teachers' capabilities to organize and use the Lego Mindstorms Ev3 as a self-efficacy variable. Teachers' self-efficacy was not significant in predicting PU; this result was not consistent with a previous study that self-efficacy supported the perceived use [19]. Teacher's self-efficacy were significant in predicting PE and BI, this found consistent with previous study [48] that Perceived ease of use was influenced by self-efficacy (the highest correlation coefficients).

In line with other research; there was a positive correlation between perceived usefulness, perceived ease of use, behavioral intention, and use in human-robot interaction in production systems [48], usefulness and ease of use were predictive of adults' attitudinal acceptance of a domestic robot in their home; ease of use and attitudinal acceptance were predictive of intentional acceptance [24], the teacher beliefs, attitudes and intention to use the software in their future teaching [49], this study found that Lego Mindstorms Ev3 as a learning resource had generated the intention of teachers to use it in their classroom so it had a positive impact on the actual use of teachers.

V. Conclusion

This study showed that the two factors of the TAM model, perceived usefulness and perceived ease to use had generated the positive attitude use of Lego Mindstorms Ev3 as learning tools by elementary school teachers, it also had a positive effect on behavior intentions, and finally, all of the variables gave support to the actual use of Lego Mindstorms Ev3 as learning tools in learning-process in elementary school.

VI. Limitation and Recommendation for Future Work

The limitations of this study: the sample size was small because it was taken from workshop participants in the introduction of robotics technology to elementary school teachers; the workshops were carried out in limited time, just 2 days, much better if we extended the duration of the workshops; the teacher was directly involved in actively interacting with Lego Mindstorms Ev3 as learning tools in the workshop but still needed adjustments at the initial stage of using Lego Mindstorms Ev3 considering that robotic technology was one of the technological trends in the 4.0 revolution era.

TAM has developed itself as a strong and robust model for explaining technical comprehension, like other theoretical frameworks. For future studies, it is necessary to develop the original model of TAM by taking into account other external variables that have an effect on the intention to use or actual use of Lego Mindstorms Ev3 in the classroom such as educational background, school facilities, gender, age, and others.

This study provides a view of teachers' acceptance of the Lego Mindstorms Ev3 in the learning process in their classroom to increase effectiveness and quality of education. This study could suggest to the elementary schools to develop learning tools based on technology, and always updating the technology will be applied in education.

Acknowledgment

This research was supported by LLDIKTI from The Ministry of Research and Technology to fund this research through the "Penelitian Strategis Nasional Institusi (PSNI)" program.

References

[1] L. Stošić, "The importance of educational technology in teaching," Int. J. Cogn. Res. Sci. Eng. Educ., vol. 3, no. 1, pp. 111–114, 2015.

[2] S. Ghavifekr and W. A. W. Rosdy, "Teaching and learning with technology: Effectiveness of ICT integration in schools," Int. J. Res. Educ. Sci., vol. 1, no. 2, pp. 175–191, 2015.

[3] B. Hendrik, N. M. Ali, and N. M. Nayan, "Robotic Technology for Figural Creativity Enhancement: Case Study on Elementary School," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 536–543, 2020.

[4] M. Masril et al., "The Effect of Lego Mindstorms as an Innovative Educational Tool to Develop Students' Creativity Skills for a Creative Society," J. Phys. Conf. Ser., vol. 1339, no. 1, 2019.

[5] S. B. Kert, M. F. Erkoç, and S. Yeni, "The effect of robotics on six graders' academic achievement, computational thinking skills and conceptual knowledge levels," Think. Ski. Creat., vol. 38, no. March, p. 100714, 2020.

[6] F. B. V. Benitti, "Exploring the educational potential of robotics in schools: A systematic review," Comput. Educ., vol. 58, no. 3, pp. 978–988, 2012.

[7] C. Oranç and A. C. Küntay, "Children's perception of social robots as a source of information across different domains of knowledge," Cogn. Dev., vol. 54, no. March, p. 100875, 2020.

[8] A. Edwards, C. Edwards, P. R. Spence, C. Harris, and A. Gambino, "Robots in the classroom: Differences in students' perceptions of credibility and learning between 'teacher as robot' and 'robot as teacher,'" Comput. Human Behav., vol. 65, pp. 627–634, 2016.

[9] R. Szczepanowski et al., "Education biases perception of social robots," Rev. Eur. Psychol. Appl., vol. 70, no. 2, 2020.

[10] M. Chevalier, F. Riedo, and F. Mondada, "How do teachers perceive educational robots in formal education ? A study based on the Thymio robot * Thymio BeeBot Finch," Ieeexplore.Ieee.Org, 2014.

[11] G. Mqawass, "Students' Perceptions and Acceptance of lego Robots in Syria," J. Interrupted Stud., vol. 1, no. 1, pp. 26–33, 2018.

[12] E. Z. F. Liu, "Early adolescents' perceptions of educational robots and learning of robotics," Br. J. Educ. Technol., vol. 41, no. 3, pp. 44–47, 2010.

[13] A. H. K. Yuen and W. W. K. Ma, "Exploring teacher acceptance of e-learning technology," Asia-Pacific J. Teach. Educ., vol. 36, no. 3, pp. 229–243, 2008.

[14] W. W. Goh, J. L. Hong, and W. Gunawan, "Exploring students' perceptions of learning management system: An empirical study based on TAM," Proc. 2013 IEEE Int. Conf. Teaching, Assess. Learn. Eng. TALE 2013, pp. 367–372, 2013.

[15] W. Daher, J. Abu-Hussein, and E. Alfahel, "Teachers' perceptions of interactive boards for teaching and learning in public and private high schools in the arab education system in israel," Int. J. Emerg. Technol. Learn., vol. 7, no. 1, pp. 10–18, 2012.

[16] C. Buabeng-Andoh, "Factors influencing teachers ' adoption and integration of information and communication technology into teaching: A review of the literature," Int. J. Educ. Dev. Using Inf. Commun. Technol., vol. 8, no. 1, pp. 136–155, 2012.

[17] C. M. Khee, G. W. Wei, and S. A. Jamaluddin, "Students' Perception towards Lecture Capture based on the Technology Acceptance Model,"

Procedia - Soc. Behav. Sci., vol. 123, pp. 461–469, 2014.

[18] L. D. Prasojo, A. Habibi, A. Mukminin, Sofyan, B. Indrayana, and K. Anwar, "Factors influencing intention to use web 2.0 in Indonesian vocational high schools," Int. J. Emerg. Technol. Learn., vol. 15, no. 5, pp. 100–118, 2020.

[19] A. Aypay, H. C. Çelik, A. Aypay, and M. Sever, "Technology acceptance in education: A study of pre-service teachers in Turkey," Turkish Online J. Educ. Technol., vol. 11, no. 4, pp. 264–272, 2012.

[20] E. Park and S. J. Kwon, "The Adoption of Teaching Assistant Robots : A Technology Acceptance Model Approach," Emerald Gr., 2016.

[21] F. George and M. Ogunniyi, "Teachers' Perceptions on the Use of ICT in a CAL Environment to Enhance the Conception of Science Concepts," Univers. J. Educ. Res., vol. 4, no. 1, pp. 151–156, 2016.

[22] M. M. A. De Graaf and S. Ben Allouch, "Exploring influencing variables for the acceptance of social robots," Rob. Auton. Syst., vol. 61, no. 12, pp. 1476–1486, 2013.

[23] D. Pal and S. Patra, "University Students' Perception of Video-Based Learning in Times of COVID-19: A TAM/TTF Perspective," Int. J. Hum. Comput. Interact., vol. 00, no. 00, pp. 1–19, 2020.

[24] N. Ezer, A. D. Fisk, and W. A. Rogers, "Attitudinal and intentional acceptance of domestic robots by younger and older adults," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5615 LNCS, no. PART 2, pp. 39–48, 2009.

[25] D. H. Shin and H. Choo, "Modeling the acceptance of socially interactive robotics: Social presence in human-robot interaction," Interact. Stud., vol. 12, no. 3, pp. 430–460, 2011.

[26] Z. Hussein, "Leading to Intention: The Role of Attitude in Relation to Technology Acceptance Model in E-Learning," Procedia Comput. Sci., vol. 105, no. December 2016, pp. 159–164, 2017.

[27] S. Alharbi and S. Drew, "Using the Technology Acceptance Model in Understanding Academics' Behavioural Intention to Use Learning Management Systems," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 1, 2014.

[28] C. Bröhl, J. Nelles, C. Brandl, A. Mertens, and C. M. Schlick, "TAM reloaded: A technology acceptance model for human-robot cooperation in production systems," Commun. Comput. Inf. Sci., vol. 617, no. March 2017, pp. 97–103, 2016.

[29] R. Alotaibi, L. Houghton, and K. Sandhu, "Factors Influencing Users' Intentions to Use Mobile Government Applications in Saudi Arabia: TAM Applicability," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 7, pp. 200–211, 2017.

[30] A. Kim, J. Han, Y. Jung, and K. Lee, "The effects of familiarity and robot gesture on user acceptance of information," ACM/IEEE Int. Conf. Human-Robot Interact., pp. 159–160, 2013.

[31] T. I. Ensign, "Elementary Educators' Attitudes about the Utility of Educational Robotics and Their Ability and Intent to Use It with Students," 2017.

[32] C. Chalmers, "Robotics and computational thinking in primary school," Int. J. Child-Computer Interact., vol. 17, 2018.

[33] N. Reich-Stiebert and E. Friederike, "Robots in the Classroom: What Teachers Think About Teaching and Learning with Education Robots," Soc. Robot., vol. 9979 LNAI, pp. 671–680, 2016.

[34] N. N. Long and B. H. Khoi, "The Intention to Study Using Zoom During the SARS-CoV-2 Pandemic," Int. J. Emerg. Technol. Learn., vol. 15, no. 21, p. 195, 2020.

[35] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "Persuasive Robots Acceptance Model (PRAM): Roles of Social Responses Within the Acceptance Model of Persuasive Robots," Int. J. Soc. Robot., vol. 12, no. 5, pp. 1075–1092, 2020.

[36] R. Nadlifatin, B. Ardiansyahmiraja, and S. F. Persada, "The measurement of university students' intention to use blended learning system through technology acceptance model (tam) and theory of planned behavior (TPB) at developed and developing regions: Lessons learned from Taiwan and Indonesia," Int. J. Emerg. Technol. Learn., vol. 15, no. 9, pp. 219–230, 2020.

[37] A. R. D. E. R. Pütten and N. Bock, "Development and Validation of the Self-Efficacy in Human-Robot-Interaction Scale (SE-HRI)," ACM Trans. Human-Robot Interact., vol. 7, no. 3, 2018.

[38] L. Huang, T. Varnado, and D. Gillan, "Exploring reflection journals and self-efficacy in robotics education," Proc. Hum. Factors Ergon. Soc., vol. 2014-Janua, pp. 1939–1943, 2014.

[39] R. Latikka, T. Turja, and A. Oksanen, "Self-efficacy and acceptance of robots," Comput. Human Behav., vol. 93, pp. 157–163, 2019.

[40] N. L. Robinson, T. N. Hicks, G. Suddrey, and D. J. Kavanagh, "The Robot Self-Efficacy Scale: Robot Self-Efficacy, Likability and Willingness to Interact Increases after a Robot-Delivered Tutorial," 29th IEEE Int. Conf. Robot Hum. Interact. Commun. RO-MAN 2020, pp. 272–277, 2020.

[41] K. G. Tileng, "Penerapan Technology Acceptance Model Pada Aplikasi Edmodo di Universitas Ciputra Surabaya Menggunakan Analisis Jalur," Juisi, vol. 01, no. 01, pp. 28–37, 2015.

[42] S. A. Filippov, A. L. Fradkov, and B. Andrievsky, Teaching of robotics and control jointly in the university and in the high school based on LEGO Mindstorms NXT, vol. 44, no. 1 PART 1. IFAC, 2011.

[43] J. Chetty, "Lego © Mindstorms : Merely a Toy or a Powerful Pedagogical Tool for Learning Computer Programming ?," Proc. 38th Australas. Comput. Sci. Conf. (ACSC 2015), no. January, pp. 27–30, 2015.

[44] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Q. Manag. Inf. Syst., vol. 13, no. 3, pp. 319–339, 1989.

[45] B. Çukurbaşı, G. Yavuz Konokman, B. Güler, and S. Kartal, "Developing The Acceptance Scale Of LEGO Robotics Instructional Practices: Validity And Reliability Studies," Bartın Üniversitesi Eğitim Fakültesi Derg., vol. 7, no. 1, pp. 191–214, 2018.

[46] C. Kim, D. Kim, J. Yuan, R. B. Hill, P. Doshi, and C. N. Thai, "Robotics to promote elementary education pre-service teachers' STEM engagement, learning, and teaching," Comput. Educ., vol. 91, pp. 14–31, 2015.

[47] J. Han and D. Conti, "The use of UTAUT and post acceptance models to investigate the attitude towards a telepresence robot in an educational setting," Robotics, vol. 9, no. 2, 2020.

[48] C. Bröhl, J. Nelles, C. Brandl, A. Mertens, and V. Nitsch, "Human–Robot Collaboration Acceptance Model: Development and Comparison for Germany, Japan, China and the USA," Int. J. Soc. Robot., vol. 11, no. 5, pp. 709–726, 2019.

[49] I. M. Santos, N. Ali, M. S. Khine, A. Hill, U. Abdelghani, and K. A. Al Qahtani, "Teacher perceptions of training and intention to use robotics," IEEE Glob. Eng. Educ. Conf. EDUCON, vol. 10-13-Apri, no. April, pp. 798–801, 2016.

APPENDIX A. QUESTIONNAIRE DETAILS USED IN THE SURVEY

Construct	Items	Item wordings	Reference
Perceived to Used	PU1	Using Lego Mindstorms Ev3 increases my productivity in classes	[44] [45]
	PU2	Using Lego Mindstorms Ev3 increases my performance in classes	
	PU3	Using Lego Mindstorms Ev3 increases my efficiency in classes	
	PU4	I think that using Lego Mindstorms Ev3 is useful for activities related to my school life	
Perceived ease to used	PE1	I find it easy to learn Lego Mindstorms Ev3 applications	[45]
	PE2	I easily teach lessons with Lego Mindstorms Ev3 practices	
	PE3	The steps that I have to take to solve any problem in Lego Mindstorms Ev3 practice are clear and comprehensible	
	PE4	I think I will easily master Lego Mindstorms Ev3 practices	
Attitude toward used	AT1	It would be fun to do Lego Mindstorms Ev3 practices in my classes	[45]
	AT2	I would enjoy doing Lego Mindstorms Ev3 practices in my classes	
	AT3	It would make me happy to do Lego Mindstorms Ev3 practices in my classes	
Behavioral Intention to Use	BI1	I want to do Lego Mindstorms Ev3 practices in my classes	[45]
	BI2	I Would like to do Lego Mindstorms Ev3 practices in my future classes	
	BI3	I will encourage my colleagues to do Lego Mindstorms Ev3 practices	
	BI4	I will include Lego Mindstorms Ev3 practices in my education and teaching career	
Self-Efficacy	SE1	I can use the Lego Mindstorms Ev3, if someone shows me how to do it first	[28] [13]
	SE2	I can use the Lego Mindstorms Ev3, if I had only the manual book for reference	
Subjective Norm	SN1	In general, the organization supports the use of the Lego Mindstorms Ev3	[28] [36]
	SN2	those people who are important to me would strongly support my using Lego Mindstorms Ev3 in my classroom	

Simulation Study on Blood Flow Mechanism of Vein in Existence of Different Thrombus Size

Nabilah Ibrahim[1]
Faculty of Electrical and Electronic Engineering
University Tun Hussein Onn Malaysia
86400 Batu Pahat, Johor, Malaysia

Muhammad Kamil Abdullah[3]
Faculty of Mechanical and Mechatronics
University Tun Hussein Onn Malaysia
86400 Batu Pahat, Johor, Malaysia

Nur Shazilah Aziz[2]
Department of Test
Venture Technocom Systems Sdn. Bhd.
Tebrau, 81100, Johor, Malaysia

Gan Hong Seng[4]
Medical Engineering Technology
University Kuala Lumpur British Malaysia Institute
53100, Gombak, Selangor, Malaysia

***Abstract*—Blood velocity is expected to be as a parameter for detecting abnormality of blood such as the existence of thrombus. Proper blood flow in veins is important to ensure effective return of deoxygenated blood to the heart. However, it is much challenging to recognize the vessel condition due to the inability to visualize the thrombus presence in the vessel. The presence of noise in the image obtained from ultrasound scanning is one of the obstructions in recognizing it. Considering the difficulty, this study aims to assess the velocity and vorticity at the vein valve region using Computational Fluid Dynamics (CFD) method. The velocity of blood and the size of valve orifice are considered important parameters in designing the vein since the stenosis and irregularities of velocity in blood vessels are known as the risk factors for thrombus formation. From the simulation, the velocity contour plot of the blood flow can be visualized clearly. The blood distribution was presented using velocity profile while the fluid particles movement was shown by the velocity vector. The low blood velocity clearly shows the low velocity region which reside at the cusps area and at the beginning of the valve leaflets. Therefore, the present study is able to visualize and evaluate the probable location of thrombus development in the blood vessel.**

Keywords*—Blood velocity profile; velocity contour plot; computational fluid dynamic (CFD); thrombus; vein valve

I. Introduction

Deep vein thrombosis (DVT) and pulmonary embolism (PE) that relatively known as venous thromboembolism is the third leading cause of cardiovascular disorder after myocardial infarction and stroke. DVT is the development or existence of blood clot or thrombus in one of the large veins deep in the body. The lower limb is the most common site of thrombosis due to relatively slow or disturbed blood flow. The existence of thrombus may block the flow of blood through the vein partially or completely. This may lead to the rupture of the thrombus which then migrates to lungs that finally becomes embolus to occlude a pulmonary artery. This condition is called PE. According to Virchow triad theory [1], at least two of the factors occurring simultaneously increase patient risk of developing DVT. The factors are stasis, vessel damage and hypercoagulability. Stasis is believed to be one of predominant of the three factors. In addition to stasis, endothelial damage also affecting the blood flow in disruption of vessel elasticity [2]. Dysfunction of endothelial increases the expression of adhesion molecules such monocytes, leukocytes, and platelets that later contribute to the abnormalities in blood flows. This is the primary factor to lead to hypercoagulability [3]. A lower extremity DVT linked to cause an estimated 50% risk of PE if not treated in a timely effective manner. It is reported that 15% to 32% of lower extremity DVT most likely to develop PE [4]. Therefore, early diagnosis of DVT is essential to prevent unnecessary deaths from PE.

Although many diagnostic tools exist to evaluate the presence of DVT such as computed tomography (CT) and magnetic resonance imaging (MRI), the use of ultrasound (US) imaging for routine DVT evaluation is superior in accuracy, cost and feasibility [5]. On top, compression ultrasound and duplex ultrasonography are both available US methods to evaluate DVT by assessing the collapsibility of vein that may rapidly performed. Other than that, the evaluation of blood flow through a vessel is an essential aspects of cardiovascular health since it is known as primary factor that contribute to the death [6]. However, such methods are not always indicating the true behavior of blood flow especially for the complex geometry and vortex formation. Normally, the blood flow can be monitored non-invasively using ultrasound Doppler or by building a phantom mimicking the artery or vein for in-vitro test [7]. Performing such experiments usually was believed to giving out accurate result if properly executed following the biological nature. However, due to the difficulty in setting up the components, several results may come to error. Thus, the outcome might not be accurately reflecting the actual flowing of blood behavior.

An alternative to constructing a physical experiment is by performing a simulation of an actual conditions for virtually problem conditions. Computational fluid dynamics (CFD) investigation have been utilized to assess particular parameters in fluid flow, for instance wall shear stress, velocity of blood flow, and pressure [8]. Thus, CFD modelling is proposed to

better understand the mechanism of underlying DVT. Previously, we have reported in [9] the CFD simulation on blood velocity and vorticity in vessel. However, the effect on the thrombus existence did not be considered. An efficient visualization of flow field can assist in further diagnosing the cause of vein disease. This study focus to investigate the effect of valve opening, and velocity to the thrombus formation on blood flow distribution in popliteal vein.

Next section shows some previous works conducted by researchers that related to computational simulation. In Section III, vein modelling is described which using computational domain. Simulation results are discussed in Section IV. Further discussion is elaborated in Section V. Finally, the conclusions and future work are presented in Section VI.

II. Previous Works

In medical education, prediction is much more important in the way to prepare for the necessary technique and aggressive motion on patients. In particular interest for relatively non-invasive technique and low cost, ultrasound techniques are mostly favorable to be used [10-11]. Those techniques may provide results on early analysis of certain disease. Nonetheless, the internal organs are unpredictable which may lead to less reproducibility results. Here, computational simulations or numerical studies for blood flow have always preferred as it is always a realistic way to simulate organ or vessel. Some studies show the findings from the simulation works. In [12], they make comparison between non-Newtonian and Newtownian models behavior of the blood to the plaque. In the study, the pulsatile flow was used interaction with lipid pool was observed. From the simulations, it is found that the non-Newtonian model shows a higher peak for most critical parameter considering the risk or benefit ratio for carotid endarectomy. The result was then increase understanding on the plaque stability. Oppositely, the study expressed the features limitation when considering only the wall shear stress and vomises stress. These parameters even though give benefit to predict the clot formation, it is insufficient to announce the flow distribution changes along the blood vessel. Therefore, the blood flow velocity, volume flow rate and shear distribution were analyzed in [13]. The work however becomes more complex when it dealing with angio-Computed Tomography data. Moreover, the study did not mention the analysis effect of blood flow velocity and vorticity on the plaque that might be considered as one of the factor that could trigger the plaque rupture. Thus, in [14], the work discussed on the possible plaque and thrombus rupture due to the interaction with blood flow. Even though the discussion proposed in detail the numerical work on blood flow velocity which also covered monocytes motion, it did not confer with the velocity vector for each minor segment along the vessel including the plaque and thrombus region. The information of velocity profile and velocity vector is essential to predict the probable location of thrombus formation in blood vessel.

III. Vein Modelling

After completing the vein drawing, the drawing will be used into another computer simulation program known as computational fluid dynamic (CFD) Ansys-CFX. Ansys-CFX is a type of software where users are allowing to test a system by simulating the fluid behavior in virtual environment. In this software, the simulation was divided into three cases. Sizes of valve aperture, velocity and sizes of thrombus will be the variable in the simulation. This is to ensure each changes in vein is fully discovered from the simulation.

A. Computational Domain

Computational domain is a platform of solving the mathematical model of the physical problem. Throughout the mathematical works, conservation of matter, momentum, and energy must be satisfied in the region of interest. Discretization technique was applied to develop approximations of the governing equations of fluid mechanics in the fluid region of interest. This discretized domain is known as grid or mesh. In this study, there is only one computational domain that will be focusing on which is venous vessel, as shown in Fig. 1.

Fig. 1 shows the structure of vessel with dimension which been drawing using SolidWorks. Fig. 1(a) indicates the length of vein, diameter, and diameter at swell area which are 10 cm, 10 mm, and 11 mm respectively. Noted that the blue arrow denotes the direction of blood flow. The length of the valve was assumed to 5 mm and the valve leaflets thickness assumed to be 1 mm. Fig. 1(b) shows the vein model with the presence of thrombus. The thrombus size used in this study is 1 mm and 3 mm where the ration of *t*:*l* is 1:4.

Fig. 2 shows the computational domain with the implemented mesh. Here, the mesh used is tetrahedral mesh where the mesh containing structured and unstructured mesh. The body is accommodated by unstructured mesh while, for the near wall region, structured mesh or known as inflation is used. The structured mesh is implemented due to the needs of critical meshing at the near wall region.

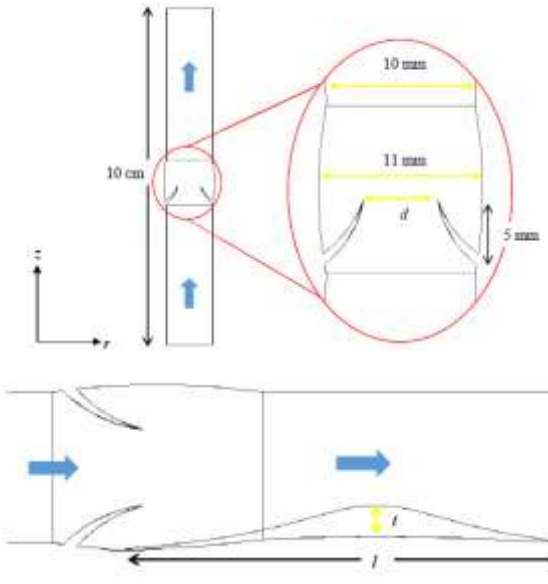


Fig. 1. Geometrical Structure of the Popliteal Vein with Dimension, a) Condition without Thrombus with *d* Represent the Size of Leaflets Opening that vary from 30% to 70%, b) DVT Condition with *t* Represent Thrombus Size and *l* Represent the Length of Thrombus.

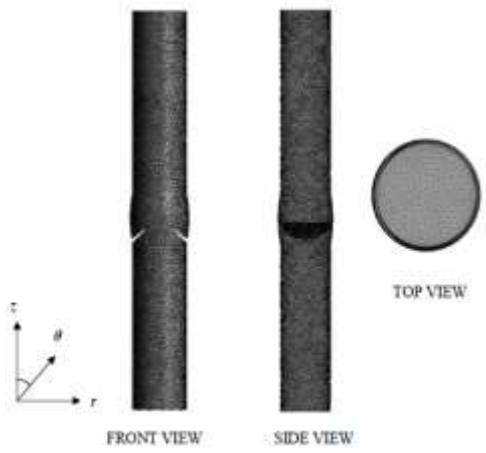


Fig. 2. Computational Domain Meshes for Front, Size and Top Views.

B. Boundary Condition and Numerical Setup

The boundary condition is to set the initial value of simulation. Here, the inlet, outlet and wall need to be assigned. Inlets are used for regions where inflow is expected while outlets are used for regions where outflow is expected.

Fig. 3 shows the specific location of the boundary applied on computational domain. The arrow denotes the direction of blood flow. Table I shows the details of boundary conditions.

The present numerical investigation was carried out with the employment of laminar model. All the numerical and boundary conditions was referred from the previous study in order to follow the real physiologic condition. Table II shows the details of boundary conditions.

C. Code of Cases

Fig. 4 shows the code uses for each type of cases. V10 represents the value of inlet velocity used. In this study, the range of velocity used was 10-50 cm/s. This is following the normal range of flow velocity in vein which actually from 10 cm/s to 40 cm/s [15]. While 20 cm/s was chosen to simulate the cases with DVT condition as to represent the normal velocity flow in resting condition [16]. O3 indicates the size of valve orifice. While NT represents case without the presence of thrombus, T1 and T3 represent the case with the presence of 1 mm and 3 mm thrombus, respectively. 1 mm size of thrombus was chosen to represent the early formation of thrombus. Meanwhile, 3 mm thrombus was to represent the severe state of DVT condition.

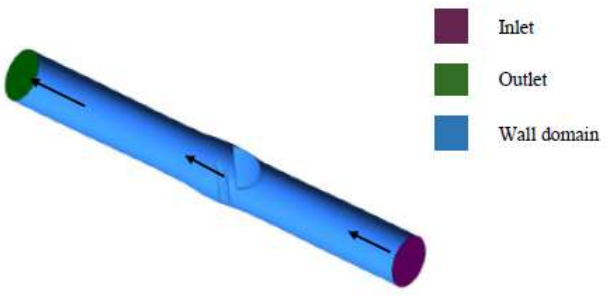


Fig. 3. Boundary Condition Location.

TABLE I. BOUNDARY CONDITION DETAILS AND FLOW PROPERTIES

Blood Viscosity (μ)	0.0035 Pa s [16, 19-20]
Blood Density (ρ)	1050 kgm^{-3} [21-24]
Temperature	37 [17,18]
Inlet (μ)	10,20,30,40,50 cms^{-1} [16]
Outlet (P)	0 mm Hg
Wall	No-slip wall condition

TABLE II. NUMERICAL SETUP

State of fluid flow	Steady
Convergence criteria	Residual type = RMS Residual target = 1.E-4 Minimum iterations = 1 Maximum iterations = 100
Model type	Laminar

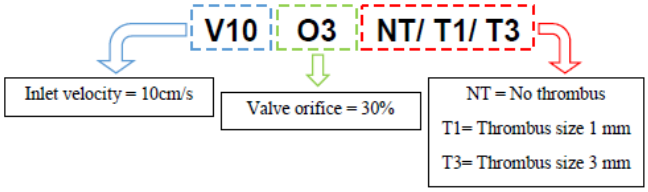


Fig. 4. Code uses for each Type of Cases.

IV. RESULTS

The results will be including the velocity contour, velocity profile, velocity vector and also vorticity of the blood flow in the vein. The variable that changes in the simulation is the size of valve orifice, velocity and thrombus size. The valve aperture was set from 3 mm to 7 mm, while velocity was set from 10 cm/s to 50 cm/s. Thrombus size used was 1 mm and 3 mm. This is to observe the blood behavior as it passes through stenosis or blockage in the blood stream.

A. Effect of Different Velocity to the Blood Flow

The information regarding blood velocity in a blood vessel might be an estimate of sufficient accuracy in many cases. Thus, this study was carried out to show the effect of blood velocity on the blood behavior at the valve region. This section was simulated to find the effect of different inlet velocity to the fluid flow inside the vessel. The different inlet velocity is representing the different velocity of human as every human carries different value of velocity insider their vessel.

Based on Fig. 5, the scaled legend of red color represents natural direction of blood flow which has the range of velocity from 5 cm/s until 95 cm/s. While blue color represents the opposite direction of blood flow which the range of velocity between 5 cm/s and -5 cm/s. It is clearly can be visualized the increasing of velocity when the fluid is passing through the valve region and keep extended to the distal side, especially for the inlet velocity from 20 cm/s until 50 cm/s. While low blood velocity region can be spotted reside in the valve cusps. The low velocity causing stagnation of blood where some particles of blood will be circling and pooling at the same place that could increase the blood contact time with the endothelial, thus will lead to formation of thrombus.

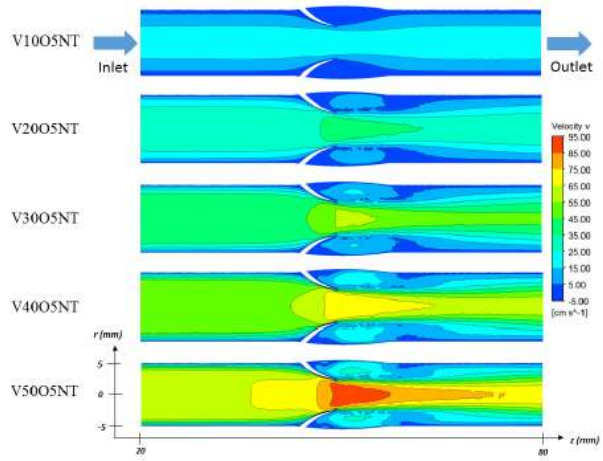


Fig. 5. The Various Value of Inlet Velocity with Valve Orifice Size 5 mm.

Fig. 6 demonstrates velocity profile for five cases with different inlet velocity. The velocity profile was divided into five section along the vessel. Two sections which A and B was plotted prior the valve while section C, D, and E was plotted after the valve. Each sections were set at 8 mm interval. In all cases, velocity profile in section A and B shows fully developed with a slight steeper increased proportionally to the inlet velocity. At the section C of V20O5NT, the velocity value started to have negative value at the vessel wall. The reverse flow can be spotted reside at the vessel wall. This condition continues until the inlet velocity reach 50 cm/s. As the velocity reach further away from the wall at section D and E, the fluid velocity changes from zero at the surface because of the no-slip condition.

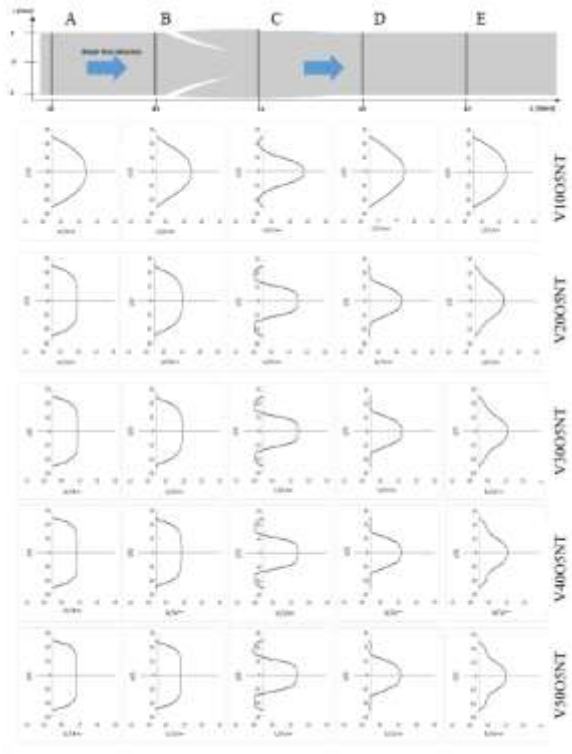


Fig. 6. Velocity Profile for different Inlet Velocity.

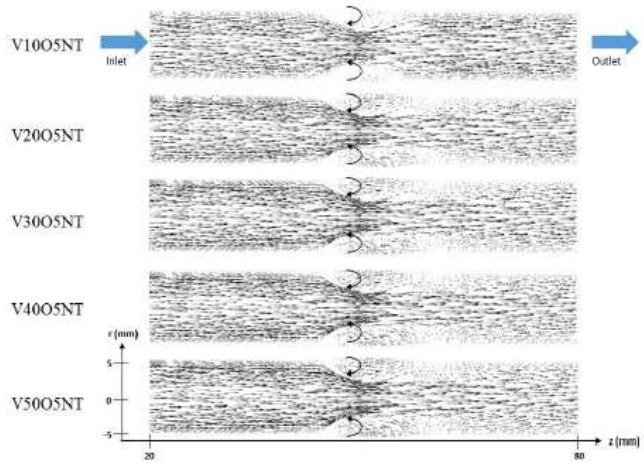


Fig. 7. Velocity Vector for each different Value of Inlet Velocity.

Fig. 7 shows the velocity vector for five cases with 50% of valve orifice along with the different inlet velocity. Based on the figure, the arrow means the flow of the stream. At the back of the valve, there are arrows which flow opposite to the z-direction which proves the back flow of blood at the cusps of the vein. The number of vector change slightly with the increasing value of inlet velocity. Here, V10O5NT shows uniform vector region along the stream. While, there are low vector region of backflow of vector at the cusps area. As velocity increasing, the vector region become stronger and the region become larger as shown by V50O5NT. The number of vector also decrease as it passing through the valve. In addition, the reversed flow region grows and expand with the increasing value of velocity. In case the velocity value is low, the condition could lead to the thrombosis development.

Vorticity contour of five cases as shown in Fig. 8 was simulated with 50% valve orifice and has different value of inlet velocity. The blue color indicates the blood particles which in the clockwise rotation, while red color indicates the anticlockwise rotation. The scaled legend showing the highest frequency value of 150 per second while the lowest vorticity frequency of -150 per second. From the all cases, it can be visualized that the vorticity near wall is slowly build up with the increasing of inlet velocity. V10O5NT shows very small region of particles moving at the valves. Meanwhile, the particles movement in V50O5NT shows the highest value of frequency due to the increasing of inlet velocity. Other than that, due to the flow separation at the valve edges, vortices were forming behind the cusps. The separation flow later reattaches at the sinus wall following the center stream at the lumen. In conclusion, low inlet velocity giving zero frequency of vorticity which most of the particles concentrating on the cusps area of vessel wall.

B. Effect of Different Valve Orifice with 1 mm Thrombus

The existence of thrombus able to completely or partially block the movement of blood particles. This section of result aiming to portray the effect of valve orifice on the thrombus formation of 1 mm in size.

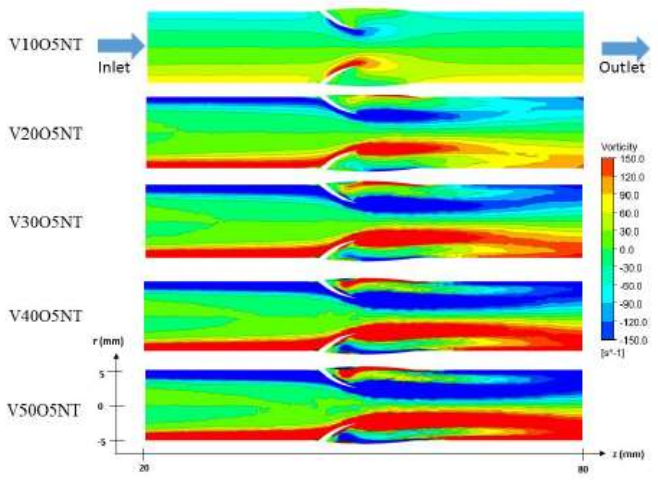


Fig. 8. Vorticity Contour Plot for five different Velocities.

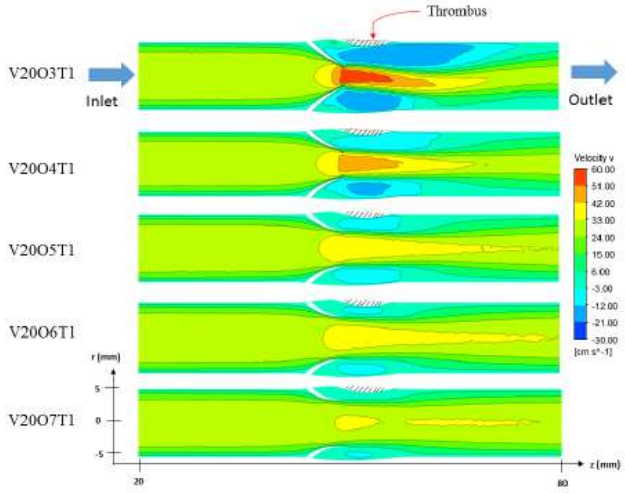


Fig. 9. Velocity Contour Plot for Five Cases with 1 mm Thrombus.

Fig. 9 demonstrate the velocity contour plot for five different valve orifice with the presence of thrombus. Each cases carried same inlet velocity which is 20 cm/s. Here, velocity value spotted to be increasing at the stenosis area. However, for case V20O3T1, the blood distribution slightly different from the normal flow. V20O3T1 shows high negative velocity region at the thrombus location. This prove that more particles are moving in opposite direction of the normal flow. Furthermore, the degree of reverse flow becomes worse with the decreasing of the valve orifice size. This indicates that the thrombus region in stenosis area causing more complex fluid flow.

Fig. 10 shows the velocity profile for each cases with different size of valve orifice. V20O3T1 showing fully developed velocity profile at the section A and B. While section C has developed negative or reverse flow zone as the fluid passing through valve area due to the sudden orifice of the vessel. Section D shows asymmetric velocity profile because of obstruction by the thrombus. The distorted velocity profile continues until section E. The same pattern at the section A and B occurs in V20O4T1, V20O5T1, V20O6T1 and V20O7T1 which showing fully develop velocity profile. At the section C and D, since the fluid just passing through the valve area, the boundary layer separation occurs where it can be observed that the fluid flow in unstable velocity value. However, the reverse flow on negative velocity value is changing with the increment of valve orifice.

Based on the Fig. 11, the positive axial velocity occupied the lumen of vessel. While the negative velocity appeared at the cusps area. The blood flowed backward as the blood occupied the cusps. V20O3T1 shows stronger velocity vector at the thrombus area since it has smaller size of orifice. Other than that, the vector size at the cusps area also decreasing with the increasing of the valve orifice. The number of vector also decrease as it passing through the valve. Here, it can be seen the reversed flow region grows and expend with the decreasing value of valve orifice.

The results in Fig. 12 show uniform pattern of vorticity contour near wall. More particles can be spotted reside at the wall area and at the edge of valve leaflets. Due to the narrowing of a vessel, high vorticity region was formed at the valve area. V20O3T1 shows slightly different pattern of vorticity as the blood passing through the thrombus area. Since it has smaller size of the orifice, the size of the vorticity region become wider compared to the other cases. It could be concluded that the vorticity region become smaller as the valve orifice become wider.

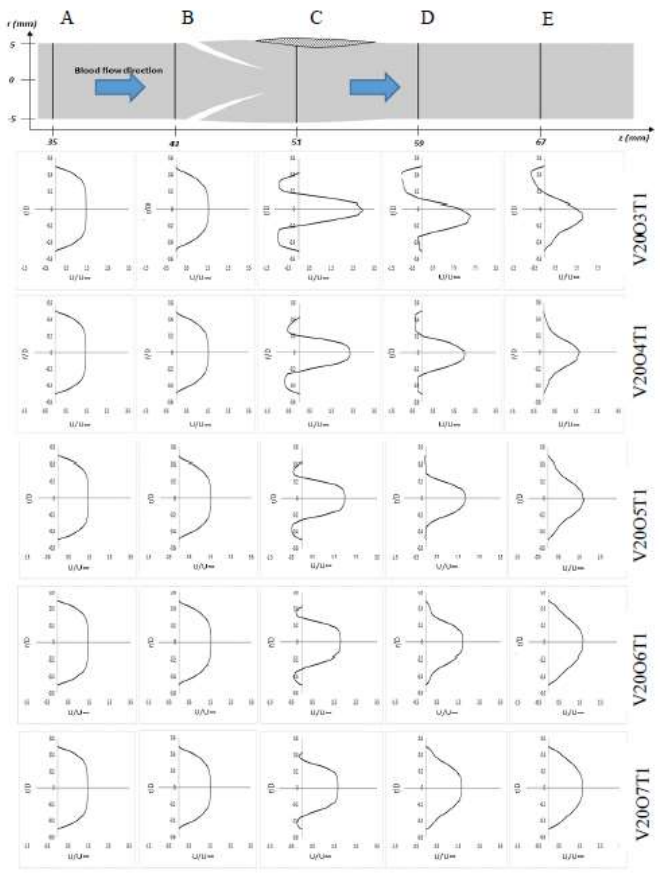


Fig. 10. Velocity Profile with 1 mm Thrombus.

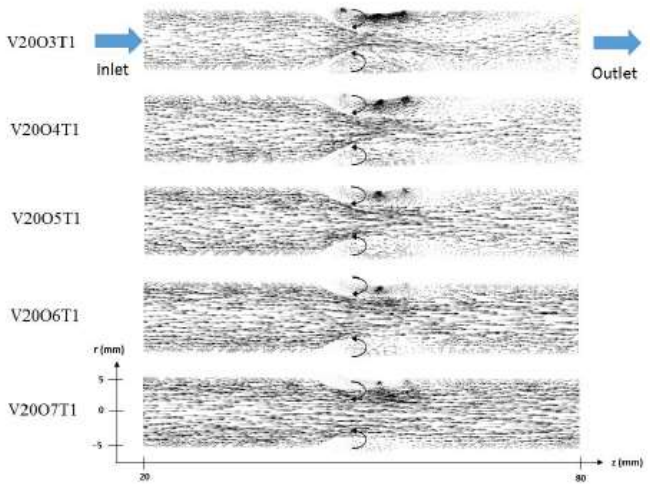


Fig. 11. Velocity Vector for Five different Valve Orifice with the Presence of 1 mm Thrombus Size.

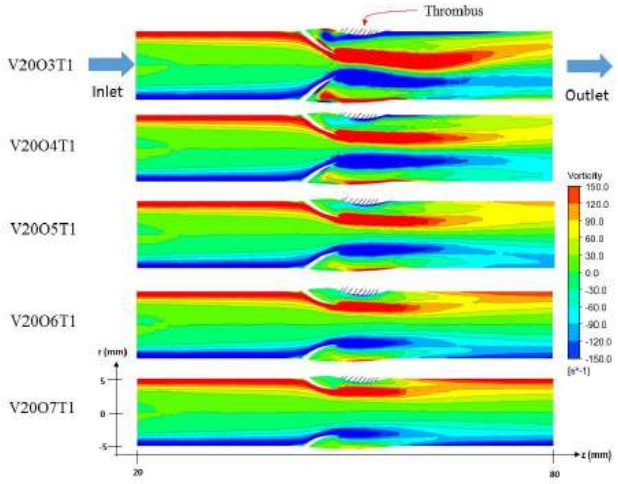


Fig. 12. Vorticity Contour Plot for Five different Valve Orifice with the Presence of 1 mm of Thrombus Size.

C. Effect of Different Valve Orifice with 3 mm Thrombus

This section was simulated to observe the effect of valve opening with the existence of 3 mm thrombus to the velocity of blood in the vein, which each cases carried the same inlet velocity of 20 cm/s.

From Fig. 13, an obvious pattern can be observed where the high velocity region only can be seen from the vessel with the smallest size of valve opening which is from V20O3T3. The high velocity region can be seen losing its momentum as the valve opening size increase. This could be concluded that blood flowing before the valve region is flow in a fully developed velocity profile. When the blood pass through the valve orifice, the velocity increase as it has to pass through the smaller area. The flow can be seen still accelerating as it passing through the thrombus area. Therefore, this shows that the increment of valve opening resulted on the decrement of the blood flow velocity in the vein. It also shows the unstable flow velocity at the thrombus area due to the disturbed flow velocity.

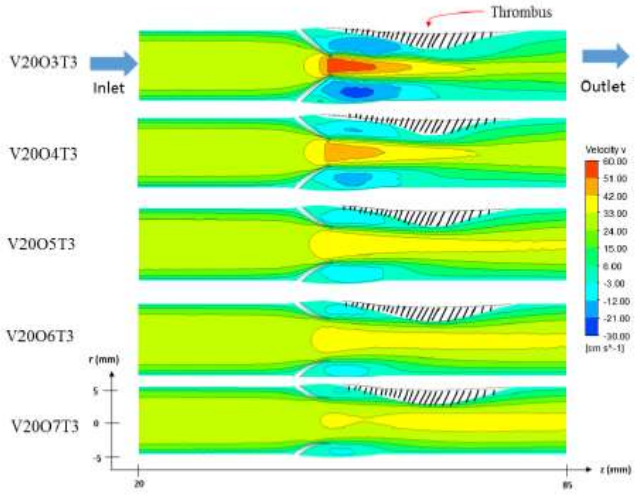


Fig. 13. Velocity Contour Plot for Five Cases with 3 mm Thrombus.

From the velocity profile shown in Fig. 14, each cases show the same pattern at section A and B with fully developed velocity profile. While section C and D has produced negative or reverse flow zone in V20O3T3 and V20O4T3 due to the obstruction by the thrombus. However, the reverse flow in section D shows decrement in the increment of valve orifice. In addition, the unstable of velocity profile can be spotted as the fluid is further away from the valve.

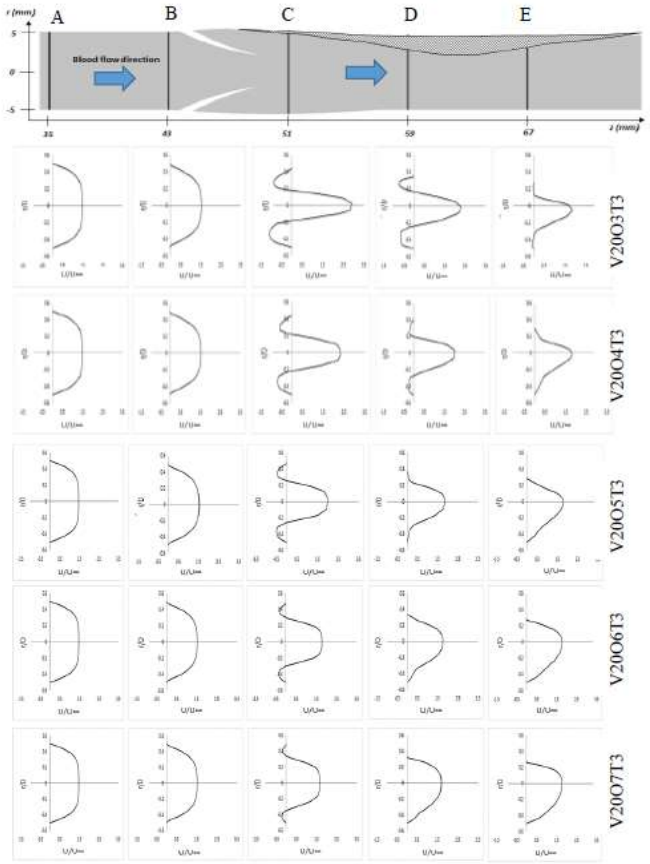


Fig. 14. Velocity Profile with 3 mm Thrombus.

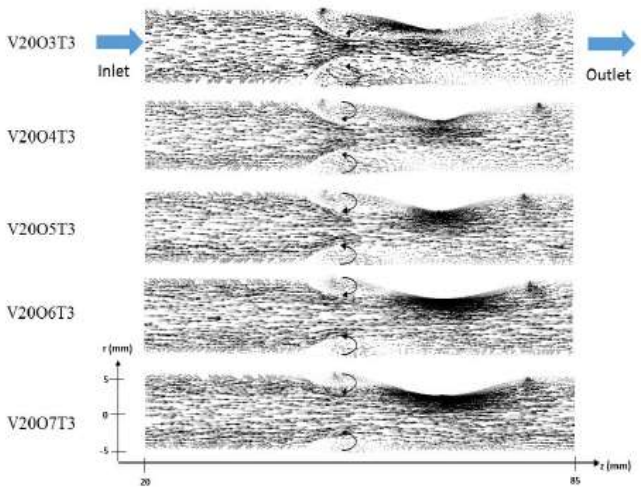


Fig. 15. Velocity Vector for five different Valve Orifice with the Presence of 3 mm Thrombus Size.

Fig. 15 shows the velocity vector for five cases with different size of valve orifice with the presence of 3 mm thrombus. The number of vector change slightly with the increasing of valve orifice. As the size of valve orifice increasing, the vortices become stronger at the thrombus area which can be seen from V20O7T3. The number of vector also decrease as it passing through the valve. Furthermore, the reversed flow region grows and expand with the decreasing value of orifice.

Fig. 16 demonstrates the vorticity contour plot of the five cases. It is clearly shows that the valve orifice size increased disproportionally to the frequency of vorticity. V20O7T3 shows very small region of particles moving at the valves area compared to the V20O3T3 where more particles can be spotted reside at the wall area and at the edge of valve leaflets.

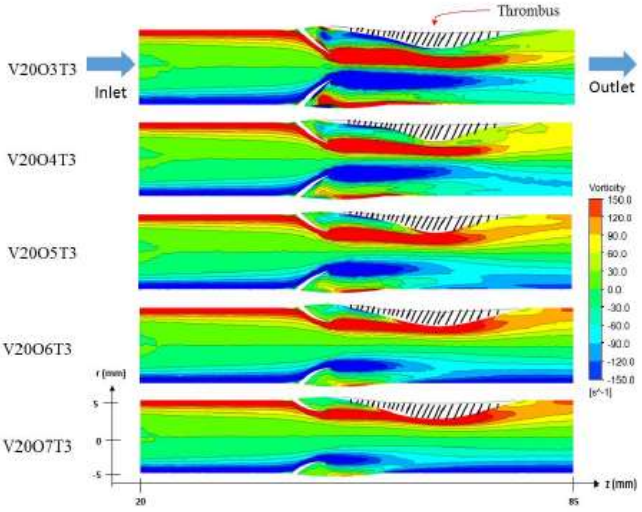


Fig. 16. Vorticity Contour Plot for Five different Valve Orifice with the Presence of 3 mm of Thrombus Size.

V. Discussions

The endothelial damage is one of the risk factor of the thrombus formation. Previous work in [1] has agreed where the vascular wall contribute to the propagation of thrombus size. Moreover, the existence of thrombus in the vessel is also assumed to be the factor of vessel damage. Formation of thrombus has been discussed as a results from clot or plaque rupture which then leading to platelet aggregation [25]. All those risk factors are consolidating along with blood flow changes in blood vessels that later affect the serious brain damage. Classically, other risk factors that implicated in thrombosis are hypertension, high cholesterol, and smoking which slightly associated with an increased risk of stroke. At the end, this study tries to predict the feasibility of some of the risk factors to the thrombus formation.

Here, it can be concluded that in all cases, blood flowing prior to the valve region flowed in a fully developed velocity profiles. When the blood pass through the valve orifice, the velocity increases as it has to pass through the smaller area. The flow can be seen still accelerating as it passing through the thrombus area and increasing proportionally to the thrombus size. Other than that, the disturbed flow velocity can be seen at the thrombus area regardless the inlet velocity value. In addition, for the velocity profile 30% and 40% of valve orifice, it shows the unstable condition of velocity due to the opening size that might be considered as narrow space for the flow to pass through. To be concluded, the existence of thrombus inside the stenotic vessel causing more complex flow and may lead to the growth of thrombus.

VI. Conclusion

The study was conducted to visualize the mechanism of blood in vein specifically popliteal vein. The computational fluid dynamics (CFD) modelling has been used to evaluate specific parameter which can affect the blood performance such as the sizes of valve orifice, inlet velocity and the presence of thrombus in the vein. In respect of CFD, the haemodynamics parameters such as velocity profile, velocity contour, velocity vector, vorticity contour was used to clearly visualize changes in blood flow as it follows the real condition of blood. The evaluation session on the vein was conducted with the presence of thrombus in two sizes that obviously obstruct the blood flow which then contribute to the thrombus alteration.

In future, it is recommended to simulate the case study using pulsatile blood flow instead of a steady blood flow. Since the wall of the vessel is assumed as rigid, it is a fact that vessel wall of a vein is collapsible. Thus, it is also suggested to study the effects of vessel wall to the blood flow. In the nutshell, present study has clearly showed the effect of blood flow velocity on thrombus development in the vessel.

Acknowledgment

Authors would like to thank the Ministry of Higher Education Malaysia for supporting this research under Fundamental Research Grant Scheme Vot. No. FRGS/1/2018/TK04/UTHM/02/24.

REFERENCES

[1] T. Esmon, "Basic mechanisms and pathogenesis of venous thrombosis." Blood Reviews, vol. 23, pp. 225-229, 2009.

[2] D. R. Kumar, E. R. Hanlin, I. Glurich, J. J. Mazza, and S. H. Yale, "Virchow's contribution to the understanding of thrombosis and cellular biology," Clin. Med. And Research, vol. 8, pp. 168-172, 2010.

[3] P. C. Bennett, S. H. Silverman, P. S. Gill, and G.Y.H, "Peripheral arterial disease and Virchow's triad," Thrombosis and Haemostasis, vol.101, pp. 1032-1040, 2009.

[4] S. Khaladkar, D. Thakkar, K. Shinde, D. Thakkar, H. Shrotri, and V. Kulkarni, "Deep vein thrombosis of the lower limbs: A retrospective analysis of doppler ultrasound findings," Med. J. of Dr Dr. D.Y. Patil University, vol. 7, pp. 612-619, 2014.

[5] B. K. Zierler, "Ultrasonography and diagnosis of venous thromboembolism," Circulation, vol. 109, pp. 1-14, 2004.

[6] R. Hajar, "Risk factors for coronary artery disease: Historical perspectives," Heart Views, vol. 18, pp. 109-114, 2017.

[7] N. Ibrahim, W. N. Wan Zakaria, N. Aziz, and M. K. Abdullah, "Construction of phantom mimic vessel for study of human vessel conditions in deep vein thrombosis," IFMBE Proceedings, vol. 46, pp. 402-404, 2015.

[8] M. Selmi, H. Belmabrouk, and A. Bajahzar, "Numerical study of the blood flow in a deformable human aorta," Appl. Sci., vol. 9, pp. 1216-1227, 2019.

[9] N. Aziz, N. Ibrahim, M. K. Abdullah, N. H. I Mat Harun, "Computational fluid dynamics simulation on blood velocity and vorticity of venous valve behaviour,". Lecture Notes in Electrical Engineering, vol. 398, pp. 617-625, 2017.

[10] Lili Niu, Ming Qian, Wei Yang, Long Meng, Yang Xiao, Kelvin K.L. Wong, Derek Abbott, Xin Liu, Hairong Zheng, "Surface roughness detection of arteries via texture analysis of ultrasound images for early diagnosis of atherosclerosis," PLOS ONE, vol. 8, pp. 1-11, 2013.

[11] Angels Betriu-Bars, Elvira Fernandez-Giraldez, "Carotid ultrasound for the early diagnosis of atherosclerosis in chronic kidney disease," Nefrologia, vol. 32, pp. 7-11, 2012.

[12] Md. Rejaul Haque, Md. Emran Hossain, A.B.M. Toufique Hasan, "Effect of non-Newtonian behaviour on fluid structural interaction for flow through a model stenosed artery," Procedia Engineering, vol. 90, pp. 358-363, 2014.

[13] Z. Tyfa, D. Obidowski, P. Reorowicz, L. Stefanczyl, J. Fortuniak, K. Jozwik, "Numerical simulations of the pulsatile blood flow in the different types of arterial fenestrations: comparable anaylsis of multiple vascular geometries," Biocybernetics and Biomedical Engineering, vol. 38, pp. 228-242, 2018.

[14] N. El Khatib, O. Kafi, A. Sequeira, S. Simakov, Yu. Vassilevski, V. Volpert, "Mathematical modelling of atherosclerosis," Math. Model Nat. Phenom, vol. 14, pp. 1-25, 2019.

[15] R. Morris and J. P. Woodcock, "Evidence-based compression. Prevention of stasis and deep vein thrombosis," Ann. Surg., vol. 239, pp. 162-171, 2004.

[16] Y. J. Zhang, T. Struffert, J. Hornegger, " Simulation of the interaction between blood flow and atherosclerosis plaque," IEEE Eng. Med. Biol. Soc., pp. 1699-1702, August 2007 [29thAnnual Conf. IEEE EMBS Cite International, p. 1699, 2007].

[17] K. Chandran, S. Rittgers, A. Yoganathan, Biofluid Mechanis. Boca Raton, CRC/Taylor&Francis, 2007.

[18] Baskurt O.K., "Pathophysiological significance of blood rheology," Turkish Journal of Medical Sciences, vol. 33, pp. 347-355, 2003.

[19] Y. Jiang, J. Zhang, W. Zhao, "Effect of the inlet conditions and blood models on accurate prediction of hemodynamics in the stented coronary arteries," AIP Advances, vol. 5, pp. 057109-1-057109-9, 2015.

[20] Vinoth R., Kumar D., Raviraj A., Vijay Shankar CS, "Non-newtonian and newtonian blood flow in human aorta," Biomedical Research, vol. 28, pp. 3194-3203, 2017.

[21] K. D. Dennis, D. F. Kallmes, and D. Dragomir-daescu, "Cerebral aneurysm blood flow simulations are sensitive to basic solver settings," J. Biomech., vol. 57, pp. 46–53, 2017.

[22] L. S. Hong, M. A. Hisham Mohd Adib, M. Uzair Matalif, M. Shafie Abdullah, N. Hartini Mohd Taib, R. Hassan, "Modelling and simulation of blood flow analysis on simplified aneurysm models," IOP Conf. Series: Materials Science and Engineering, vol. 917, pp. 1-10, 2020.

[23] Yue Zhou, Chunhian Lee, Jingying Wang, "The computational fluid dynamics analyses on hemodynamic characteristics in stenosed arterial model," J. Health Eng., vol. 2018, pp. 1-6, 2018.

[24] Ana Paul N., Fernando Silva M., Frederic P., Alberto M., Ignacio L., Jean-Marc G., Cecile p., Charles A. S., Ayache B., "A clinically aligned experimental approach for quantitative characterization of patient-specific cardiovascular model," AIP Advances, vol. 10, pp. 045106-1-045196-10, 2020.

[25] M. Koupenova, B. E. Kehrel, H. A. Corkrey, J. E. Freedman, "Thrombosis and platelets: an update," European Heart Journal, vol. 38, pp. 785-791, 2017.

Singer Gender Classification using Feature-based and Spectrograms with Deep Convolutional Neural Network

Mukkamala S.N.V. Jitendra[1], Dr. Y. Radhika[2]
Department of Computer Science and Engineering, GIT
GITAM (Deemed-to-be University), Visakhapatnam-530045, AP, India

***Abstract*—The task of music information retrieval (MIR) is gaining much importance since the digital cloud is growing sparklingly. An important attribute of MIR is the singer-id, which helps effectively during the recommendation process. It is highly difficult to identify a singer in the case of music as the number of signers available in the digital cloud is high. The process of identifying the gender of a singer may simplify the task of singer identification and also helps with the recommendation. Hence, an effort has been made to detect the gender information of a singer. Two different datasets have been considered. Of which, one is collected from Indian cine industries having 20 different singer details of four regional languages. The other dataset is standard Artist20. Various spectral, temporal, and pitch related features have been used to obtain better accuracy. The features considered for this task are Mel-frequency cepstral coefficients (MFCCs), pitch, velocity, and acceleration of MFCCs. The experimentation has been done on various combinations of the mentioned features with the support of artificial neural networks (ANNs) and random forest (RF). Further, the genetic algorithm-based feature selection (GAFS) has been used to select the suitable features out of the best combination obtained. Moreover, we have also utilized the recent popular convolutional neural networks (CNNs) with the support of spectrograms to obtain better accuracy over the traditional feature vector. Average accuracy of 91.70% is obtained for both the Indian and Western clips, which is an improved accuracy of 3% over hand engineering features.**

Keywords*—Gender identification; spectrogram; genetic algorithm-based feature selection (GAFS); music information retrieval (MIR); music recommendation; and singer's gender identification

I. Introduction

Technological advancements in the music industry have created an enormous number of music clips. It is difficult to categorize and organize such several clips if proper meta-information is not provided [1]. Hence, it is essential to provide the meta-information for the available clips of the digital cloud. Moreover, it is impractical to provide the meta-information for millions of tracks available in the digital cloud. The meta-information could be related to artists, instruments, genre, lyrics, etc. Of which, artist information is a much important factor where a majority of the listeners are usually listening to the songs of their favorite artist [2]. The artist is further characterized by a singer, an artist, or a composer. There is a small difference between an artist and a singer. A singer who contributes his vocals to the portion of a song during studio recordings. An artist who performs his skills on a stage in front of the audience. In general, a majority of the audience shows interest in the songs of a particular singer. For instance, Shreya Goshal is one such Bollywood singer who gets the attraction of most Indian listeners [3].

Since it is impractical to provide the meta-information manually due to the availability of million numbers of tracks, there should be an alternative approach for the provision of meta-information. The process of automatically extracting the information from music clips is called music information retrieval (MIR). Hence, there is a good amount of research happening to investigate an alternative approach or automatic approach which extracts the meta-information. The research on MIR is initiated during the initial years of the 21st century [4].

There are many works such as singer identification, genre classification, singer identification, lyrics transcription, instrument identification, mood estimation, and music annotation done on the aspects of MIR. However, the application which is designed for a particular regional song is not giving the same performance as the songs of other regions. Hence, a challenging event, called music information retrieval evaluation exchange (MIREX) has been initiated under the international music information retrieval systems evaluation laboratory (IMIRSEL) in the year 2000.

There are thousands of papers that have been published in ISMIR since 2000 on various MIR works mentioned [1]. This would give a clue to understanding the importance of automating the MIR tasks. From the above-mentioned tasks of MIR, we have identified one important problem named gender classification which is a sub-problem of singer identification. There are around 50,000 singers in the world. There are certain singers' datasets available with 3,000 and 48,800 singers. They are provided by the institute of computational perception of Johannes Kelper university, Linz, Austria, and are called c3ka and c49ka [5]. The process of singer identification gets complicated when there is an increase in the number of singers. Hence, it is essential to further categorize the singers based on their characteristics. One such important characteristic is gender. Based on its importance for MIR, the same has been considered for this work. In general, the task of singer gender identification is named automatic singer gender identification (ASGID).

In the case of speech recognition, the task of automatic gender identification (AGID) has been used in speaker recognition, biometric systems, security, and surveillance. Since the main objective of biometric systems is to identify a person, it was mentioned that the task of gender identification simplifies the task by segmenting a person into either male or female category [6]. Moreover, the implementation of AGID is also helpful in assigning a male customer care agent to the male customer in the case of transferring calls to an agent. And in some automatic regional language identification [7]. Similarly, some research has happened to categorize the genders of singers, which simplifies the process of singer identification. This task is further helpful for the music recommender system as well while recommending the songs to the listeners [8].

However, a majority of the research has happened in detecting the gender of a speaker. A less focus has been done on the case of the gender identity of a singer. However, the features related to speech processing are found to be sufficient to model the music as well. Hence, the features that can discriminate against gender have been considered and experimented within this work. The features such as Mel-frequency cepstral coefficients (MFCCs), variations of MFCCs such as velocity and acceleration, features related to pitch have been considered for this work. Artificial neural networks and random forests have been considered as classifiers to classify the category of feature dimension [47]. The dimensionality of the feature vector is 43. Feature selection methods such as principal component analysis (PCA), and cross-correlation analysis (CCA) are considered to select suitable features from the larger dimensionality. Further, the results have been compared with the popular convolutional neural networks (CNNs) by feeding the spectrogram images of the audio signals.

The rest of the paper is laid as follows: Section II gives detailed literature done for the speech and audio signals. Section III proposed methodology with a flow diagram and the details of features is described with different classification, models were elaborated along with the implementation of spectrogram-based CNN. Section IV presents the results and implications of work with the comparisons that are given for traditional feature-based approaches and recent popular convolutional neural networks. Section V concludes the work with future directions.

II. Background

The task of automatic singer gender identification (ASGID) is essential in simplifying the process of singer identification. It also helps in indexing and categorizing the audio clips into a class of male and female singers. Hence, it can be considered as one important factor while recommending songs to the listeners. However, the research which is done on gender identification is not up to the mark. The reason could be the similarities that can be identified among male and female singers. They may expose similar characteristics while singing a song. Whereas, one can observe the differences in terms of the pitch in the case of speech [9]. Based on this, we can conclude that the task of gender identification is a challenging issue when compared to gender identification for speech. Here, we have provided the literature on both the aspects of speech and music processing. It gives a clear understanding of the similarities and differences in both aspects.

The features considered for speech and music processing are divided into three classes namely low-level, mid-level, and high-level features. A signal of shorter length which is ranging from 10 to 100 milliseconds. It gives low-level inherent information to the researcher and also provides a way to map every portion of the signal with relevant information. Further, information from larger frames −frames is nothing but a portion of signal of the same length− has been considered to extract the mid-level information. To avoid loss of any information of the signal, a technique called overlapping has been introduced and generally, 50% of the frame will be considered to overlap [10]. However, the features extracted from low-level and mid-level are useful to provide high-level information i.e. gender, singer, artist, genre, raga, etc. This high-level information is useful to recommend or categorize the audio clips.

Fundamental frequency (F0) is one important feature in recognizing gender information [11]. The pitch range is around 100 Hz to 200 Hz in the case of males, and the same is around 120 Hz to 350 Hz for females [12]. Hence, the pitch has been used as a primary feature in most of the applications that are designed to categorize gender [13]. The accurate estimation of the fundamental frequency (F0) has been used to compute the set of acoustic features that are further used in various research works to estimate the gender of a speaker/singer [14]. However, the process of estimating the accurate F0 itself is a challenging task. An algorithm is yet to be designed to compute the accurate F0 value. The reason for poor performance obtained with the gender classification systems is due to the imprecise F0 obtained with the existing algorithms. Henceforth, various other spectral features obtained from the frequency spectrum have been used as supporting features for F0 to improve the performance of gender classification systems. Some notable features are including linear predictive coefficients (LPCs), Mel frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), pitch class profile (PCPs), perceptual linear predictive coefficients (PLPs), relative spectral MFCCs (RASTA-MFCCs), relative spectral PLP coefficients (RASTA-PLPs), etc. In contrast, some research works have concluded that the traditional features that are used for the tasks of speech recognition may not be suitable for the gender classification task. Further, a scope has been found to investigate the characteristics of the signal for different genders, which gives a clue to compute the relevant features for gender-specific tasks [15]. It could be possible to obtain better accuracy if the relevant features for gender are explored.

Humans have been categorized into male or female based on certain characteristics. Speech is one such important factor that helps in recognizing the male or female. The physical parameter of the glottis, vocal tract length, and thickness decide the category of a person to the male or female class. They are generally called acoustic parameters. Several works have been initiated to recognize gender with a variety of features related to acoustic parameters and popular classification models. A variety of performances were

observed with the features identified through acoustical parameters. Of which, pitch and first formant are the prominent features found in many research work with improved performance. Pitch is the feature which is related to the source of the voice and the first formant (F1) is related to vocal tract information. An approach of linear predictive analysis has been considered to compute the pitch and F1. Certain analysis has given the information that the pitch and F1 values of males are less when compared to those of females. Distance measure has been considered based on Euclidean distance to segregate the genders using the nearest neighbor classifier. Further, it is found that the features based on autocorrelation, cepstral analysis, linear prediction, vowels that are extracted from speech, reflection in voice, fricatives did well to identify the gender of a speaker. Pitch has been considered as a primary feature. In addition to pitch, MFCC features, and energy has been supplied to support vector machine (SVM) which is resulting in the performance with an accuracy of 95% from their dataset [16]. However, pitch alone could give an accuracy of 96% with the support of neural networks. The dataset also involves information that is phoneme and speaker-independent. Moreover, various vocal source parameters are extracted to detect the gender and an average accuracy of 95.1% is obtained in detecting the male and female classes.

In some works, only the portions of voiced segments have been detected to effectively estimate the category of gender. Various cepstral features such as PLPs, LPCCs, and MFCCs have been computed from those portions to classify the gender. Also, independent dimensions of the features mentioned above have been analyzed that helps in identifying the suitable feature vector for gender classification from the category of cepstral features.

However, a majority of the works mentioned above are designed to detect the gender for the speech that is recorded in the acoustically controlled environment. Since the biometric systems can be designed effectively to recognize the person's gender, the systems that are designed for studio-recorded speech are sufficient. They may not give the required performance in the real-time environment. Moreover, the singer's voice is always accompanied by background music. In such a case, the gender detection systems designed for the studio-recorded environment may not be useful. Hence, for the first time, the gender recognition system has been designed for a noisy environment. Besides, it is also important to note that the process of gender detection is to be done for various languages. There could be a variety of parameters related to the vocal tract that may affect the performance of gender detection. Hence, the system which is designed for one language may not give a similar performance with the other languages. The author has taken care of designing an effective system that could handle various languages [17]. An accuracy obtained with this approach is 95%, which shows the capability of the system. The features related to pitch and suitable spectral features have been used to obtain accuracy.

It is also an important aspect to know about the gender detection system, which could be affected by the age factor. It is stated in the literature that the person's voice gets changed every two years due to the change in vocal tract parameters. Moreover, the gender detection system is ineffective in the case of children when compared to elders. As there could not be many differences in the vocal tract parameters in the case of children, both male and female voices look similar. Research work has been done to verify the same by [18]. Various recordings have been collected from the age groups of 8-10 years and 16-20 years. Further, experimentation has been done to detect the genders of these two groups. An accuracy of 60% and 95% have been obtained from the children identification (CID) Dataset. Features that are based on acoustical cues, prominent peaks from cepstrum, pitch based on harmonic-to-noise ratio, and source spectral magnitude have been considered for this work. Similar experimentation has been done to compare the performance of the systems with two different databases [19]. A total set of features contains the vector size with 113 dimensions. A naive Bayes classifier has been used to classify the data into the male and female categories. However, the system is not giving an accurate performance in detecting the gender of a person.

Modified voice contour (MVC) is used to measure the intensity in voice in the speech sample, which further helps in discriminating against the genders [20]. The dataset has been collected, which is forming the signal with Arabic digits. As the standard dataset for the gender recognition system is the TIMIT dataset, the experimentation setup was compared with TIMIT using the supervised support vector machine (SVM) classifier [21].

A different approach to the classification of genders has been proposed in this work. Since the proposed work is mainly focusing on gender detection for singers, it is essential to develop a system that could handle the background music as well. It is not possible to find a song without background accompaniment. Moreover, 99% of the song portions are accompanied by instrumentals. Hence, we made an effort to suppress the noise up to some extent using Chebyshev infinite impulse response (IIR) filters. However, we are unable to neglect all the background support with this approach. Therefore, it is decided that the proposed system should be effective even if the background accompaniment is there. The database is collected from the Indian Bollywood, Tollywood, Kollywood, and Sandalwood cine industries that are involving four regional languages of India, namely Hindi, Telugu, Tamil, and Kannada. As there is no standard dataset for Indian language speaker's gender detection, we could not compare the proposed approach with any other work. However, we made an effort to compare the proposed work with the Western popular artist dataset called Artist20. Various features such as MFCCs, velocity, and acceleration of MFCCs called ΔMFCCs, and Δ ΔMFCCs have been computed. Besides, pitch related features are also computed to add strength to the above-mentioned spectral features. These features are fed to the two popular non-linear classifiers, such as random forest (RF) and artificial neural networks (ANNs). Further, we have used the genetic algorithm-based feature selection (GAFS) algorithm proposed by Murthy et al. to reduce the dimensionality and complexity issues. Moreover, the popular convolutional neural networks (CNNs) are also used by feeding the spectrograms as images. It is found that CNN's are more capable of discriminating against the gender of singers.

A. The main Focus of the Article

- Identifying suitable music data set for different regional languages in Indian and Western songs to identify singer gender.
- Implementing various feature extraction processes with GAFS (Genetic Algorithm based Feature Selection) by combining with traditional features like Mel-frequency cepstral coefficients (MFCCs), Pitch, and Temporal.
- Design a novel convolutional neural network model based on spectrogram images to Automatic gender identification of a singer in a given music track.

III. Proposed Methodology

The flow diagram of the proposed gender classification system has been depicted in Fig. 1. It has various blocks, namely dataset collection, the process of dividing the dataset into training and test sets, feature extraction, classification models, spectrogram generation, and convolutional neural networks (CNNs). This section describes the process of feature extraction, classification models, spectrogram generation, and CNNs.

B. Feature Extraction

Features represent the prominent information that is useful to discriminate different classes depending on the problem chosen. The features that are chosen for one task may not be suitable for any other task. Hence, we should perform a lot of analysis while choosing the feature vector. Since the features are to be extracted from the samples of the speech signal, we have used various signal processing approaches to identify the suitable features for the task of the singer's gender identification. Correlation is one important property that gives the similarity behavior of a particular feature over different classes. Hence, we have used the same metric to check the suitability of a particular feature. However, the evolutionary-based strategy has been used to select the relevant and effective feature vector for gender detection [22]. To select the suitable feature vector from the large dimensions to reduce the complexity we used a genetic algorithm-based feature selection (GAFS) algorithm and features related to voice source and spectral are more suitable for gender detection. Hence, Mel-frequency cepstral coefficients (MFCCs) and pitch related features such as minimum pitch (P_{min}), maximum pitch (P_{max}), average pitch (P_{avg}), and deviation in pitch (P_{std}) have been considered as base features for the task of the singer's gender identification.

Also, the variations in MFCCs such as velocity and acceleration called ΔMFCCs and ΔΔMFCCs have been computed by finding the first-order and second-order differentiation on MFCCs. MFCCs (13), ΔMFCCs (13), ΔΔMFCCs (13), and pitch (4) are together forming a feature vector of length 43. The process of computing features has been detailed below:

1) Mel-frequency cepstral coefficients: One of the prominent features in extracting relevant information from the speech signal is the MFCC feature vector. It can effectively model the music signal as well. Hence, it has been used as a base-line feature in many applications such as vocal and non-vocal segmentation [23], singer identification [24], genre recognition [25], etc., that are related to the music signal. It is a representation of the short-time power spectrum and computed using the non-linear Mel scale [26]. The log magnitude of the power spectrum has been computed to construct a cepstrum. The non-linear Mel scale will be applied to extract the prominent peaks after applying the triangular band filters on the cepstrum. However, MFCCs are extracted from the short-time frames and hence, come under the category of low-level features. The signal is divided into chunks of the length 25 ms with an overlap of 10 ms. The features have been computed from the 25 ms lengthened frames [27]. The steps to compute the MFCCs are given below:

- Divide the signal into the chunks of frames of length 25 ms with an overlap of 10 ms.
- Construct a spectrum using fast Fourier transformation (FFT).
- Construct a power spectrum from the FFT computed and compute the log magnitude of the spectrum which gives a cepstrum.
- Map the powers of the spectrum obtained above onto the Mel scale, using triangular bands.
- Identify the logs of powers at each Mel frequencies.
- Take the discrete cosine transform of the list of Mel log powers, which gives MFCCs.
- Consider the prominent 13 to 39 peaks and ignore the rest. Here, the first 13 MFCCs have been considered for experimentation.

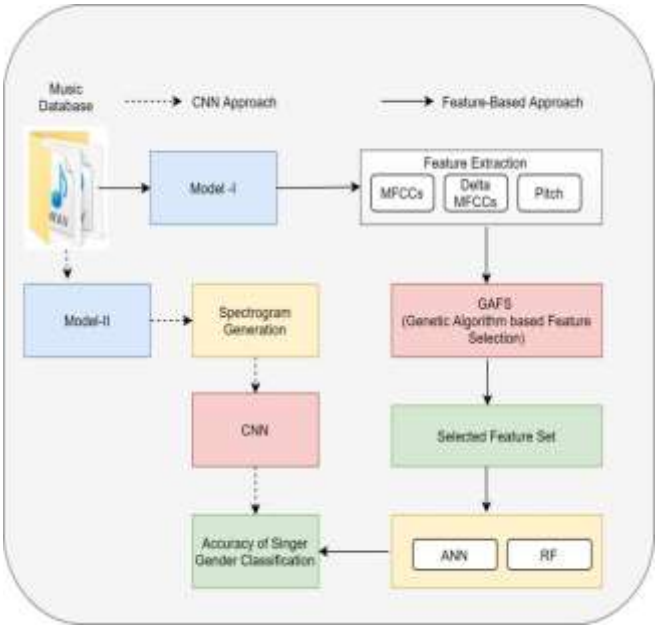


Fig. 1. Proposed Flow Work for Singer's Gender Detection and Comparison with Feature and CNN Approaches.

2) Velocity and Acceleration of MFCCs: The dynamic features that can extract the aggressive behavior of the music signal are temporal features. One such temporal feature is velocity and acceleration that are obtained from the static MFCCs. As MFCCs are static, they may not be able to detect the temporal information of the signal. The temporal information may be used in detecting the gender information of a singer. As the singers change their pitch information intentionally, the temporal information may give some clue in identifying the gender information. Hence, velocity and acceleration features have been used as supportive features for MFCCs. We found proper discrimination in the classification performance after adding velocity and acceleration features. The velocity features are extracted by computing first-order derivatives on the MFCCs and the acceleration features are the second-order derivatives [28-30]. Hence, they are generally called Delta (Δ) coefficients. Where the velocity features are represented as Δ , acceleration as Δ Δ of MFCCs. The general formula to compute the Δ coefficients are given in Eq. (1).

$$\Delta \mathrm{Cep}_{\mathrm{k}}(\mathrm{t}) = \frac{\sum_{\mathrm{i}=-\mathrm{n}}^{\mathrm{n}} \mathrm{m_i Cep_k(t+1)}}{\sum_{\mathrm{i}=-\mathrm{n}}^{\mathrm{n}} |\mathrm{i}|} \quad (1)$$

Where $Cep_k\ (t)$ is the MFCC feature that represents the k^{th} feature at time frame *t.* The total number of successive and predecessor frames are denoted with *n* and the weight mi has been added to the i^{th} frame. In general, the value of *n* considered for experimentation is 2. Further, the same process is applied to ΔMFCCs to compute accelerative features.

3) Pitch based Features: Since it is already mentioned that voice source parameters give much prominent information to recognize the gender of a singer effectively. Pitch is one such feature that is useful to detect gender. It has outperformed in recognizing gender from the speech. As there could be overlapping differences in the case of males and females, it would help efficiently in the case of the speaker's gender recognition. The same performance could not be observed with the pitch feature alone in the case of the singer's gender detection. Both the male and female singers can tune their pitch according to the tune of the song. This could be the main reason for an ineffective performance with the pitch feature alone. However, the temporal information observed from pitch may give some useful information to detect the gender effectively. Hence, some statistical methods are applied to pitch values for obtaining temporal information out of it. It results in obtaining the four different features such as minimum pitch (Pmin), maximum pitch (Pmax), average pitch (Pavg), and standard deviation of pitch (Pstd). We have used a harmonic-to-sub harmonic approach to obtain the pitch values as it performs well in the case of background accompaniment [31, 45].

4) Genetic algorithm based feature selection: Genetic Algorithms are designed to optimize the process to select the best solution and to discard the rest. The algorithm generates random values that are used to generate the population [32]. Initially, the random value of length 43 bits is generated. It is used to select the set of features from a total of 43 features. The population is the series of 0's and 1's, where '1' represents the feature consideration, and '0' represents its absence in the final set. The generation of the bits for the population is done through a random process [44, 46].

The selected features are then fed into the ANN and RF classifiers to get the accuracies. The accuracy obtained through this approach is highly efficient than that of the original feature set. The fitness of the population is calculated based on accuracy and the number of features selected. Higher the fitness, the greater the efficiency in the classification. This process is repeated for several epochs, and finally, an optimized set of the population is obtained. The mutation operation is performed by inverting or changing the bit values in the population. This process enriches the qualities of the child.

C. Spectrogram based CNNs

It is highly difficult to extract suitable features from the music signal as it is always accompanied by background music. It is also a known fact that the recent popular convolutional neural networks (CNNs) are doing well to classify the highly non-linear data. They already outperform in the field of image processing. Hence, we utilize the same for implementing the task of gender classification of Indian singers. As images are the possible input that we can feed to CNN to compute the suitable features automatically, spectrograms are constructed. Spectrograms represent the three-dimensional view of the signal having time, frequency, and intensity as the x, y, and z planes, respectively. The details of the spectrograms and the components of CNN are given in this section.

1) Spectrogram generation: Spectrograms help in analyzing the time-frequency information effectively. The frequency modulation can be observed in the case of spectrograms where it is not possible in the time domain. In general, the frequency domain gives information concerning single-frequency components. Time-frequency distribution (TFD) resolves the issue by providing both time and frequency information. Spectrograms give the information related to the moving sequence of the local spectra to any music signal [33, 34]. There are several ways of computing spectrograms. In this work, a short-time Fourier transformation has been utilized to construct a spectrogram. As the process isolates the distinguished components of two gender classes, the signal f*(t)* has been multiplied with the succeeding time windows which were shown in Eq. (2).

$$\mathrm{f(t)} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{f(t)}\, \omega(\mathrm{t} - \tau_{\mathrm{m}}) \quad (2)$$

The computed spectrograms have been used as RGB images having three dimensions for convolutional neural networks. All the images are scaled to the size of 128 * 128 pixels based on the normalization strategy.

D. Convolutional Neural Networks (CNNs)

One possible way to automatically extract the features from the images is by using the convolution operation. Various convolution layers are to be used to obtain the relevant features from the spectrogram images. With this nature, convolutional neural networks (CNNs) became popular in recent years. A majority of the applications that are based on image processing have been redesigned with the involvement of CNN. It is found that CNN's are outperforming in most of the cases when compared to traditional feature-based classification models. It is not possible to say that the CNNs are not based on a feature-based approach. However, they could extract suitable features automatically.

It is also a known fact that the speech signal is a one-dimensional time-invariant signal which gives no clue to estimate what it contains. Moreover, a music signal is highly complex when compared to the speech where instrumentals always accompany it. Hence, the task of the gender classification of a singer is harder than a speech signal. The features that are computed from time-domain and frequency-domain may give some base-line performance to discriminate against the gender of a singer. Hence, a three-dimensional spectrogram has been utilized as an image. It contains information related to frequency components and their intensity [35, 36]. It may further help in accurately classifying the gender of a singer. The components of CNNs have been detailed below, which gives fundamental information about the procedure.

A deep learning algorithm that could take an image as an input is the convolutional neural network (CNN), also called ConvNet. Various objects of the image are getting importance with the ConvNet and hence, able to discriminate each portion of the image using them. As it is known that the traditional feature-based approaches involve complicated preprocessing before extracting certain features from the image. However, CNN's can reduce the difficulty involved with the traditional approach. We have to manually change the filters to obtain the relevant features in the case of a feature-based approach. However, CNN has the inbuilt ability to apply the possible number of filters automatically.

The basic architecture of CNN for the problem chosen is given in Fig. 2. The organization of the human brain is the inspiration for designing such a connecting pattern of ConvNets. One more important piece of information that has to be noted for CNNs. The CNN might give an average performance with the grayscale images when compared to the RGB images. Hence, color spectrograms have been utilized for this task instead of monochrome images. Features obtained based on hand engineering may not be able to capture spatial and temporal variations from the spectrogram. However, ConvNets can estimate them effectively to extract suitable features. Moreover, the number of features selected is less and relevant when compared to the traditional method. The parts of CNN include the convolution layer(s) (CONV-RELU), pooling layer(s) (POOLING), fully connected network (FCN), and softmax layer (SOFT).

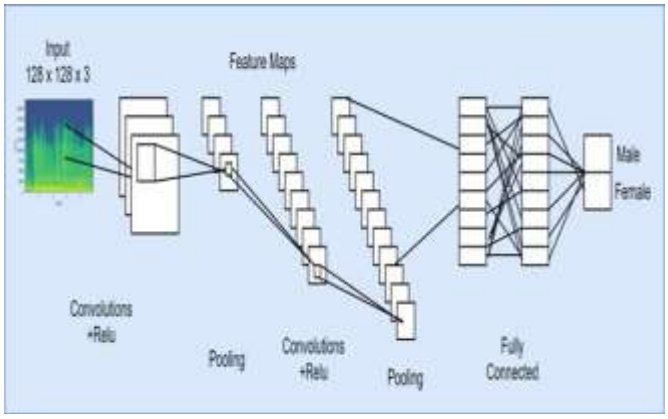


Fig. 2. The Components of the Convolutional Neural Network from Feeding Images to the Identification of Class Labels.

Algorithm for Automatic Singer Gender Identification (ASGID)

Input: *Spectrogram Images (Indian and Western Songs)*
Output: *Classification of Singer gender (Male or Female)*

1. *Start*
2. *Take* ***n * n * 3*** *images, the filter of size* ***f * f****. Were, n = 128, and f = 8 for the CNN0.*
3. *Padding* ***p = 2*** *to get effective results with the CNNs.*
4. *Generate Spectrogram image with the size of* ***(n+ 2p−f+1) * (n+2p−f+1) * f.***
5. *Consider max pool with a filter size of* ***2 * 2*** *and labeled the max filter as* ***w * w****.* ***Stride (s) = 2****.*
6. *The outcome of the pooling layer with* ***n1, w, and s*** *is* ***(n1 - f)/s + 1 * (n1 − f)/s + 1.***
7. ***Repeat*** *from Step 2* ***Until k!=4***
8. *Generate optimal features from the input spectrogram.*
9. *Flattening the output vectors of* ***Three channels into 2 classes*** *(Male or Female)*
10. *Stop*

IV. Result and Discussions

This section mainly focuses on two aspects. One is the datasets that are used for the task of gender classification, and the other part gives the detailed observations of the results obtained using the proposed approach.

A. Dataset Collection

Two datasets have been used for this work. One is the Indian popular songs dataset (IPSD), which has been designed with 20 singers. The other dataset is the standard Artist20 dataset. The IPSD has been designed with 250 audio clips. The average length of the audio clip is five seconds. The dataset includes song clips of various Indian cine industries, including Bollywood, Kollywood, Sandalwood, and Tollywood. The clips are based on regional languages, namely Hindi, Tamil, Kannada, and Telugu, respectively.

Further, each clip is segmented into 25 ms segments and an overlap of 10 ms. All the clips are recorded at the sampling frequency of 44100 Hz. The dataset has been collected based on the study done in [22]. Care has been taken to involve various background accompaniment instruments while collecting the dataset. The reason for selecting the songs of various languages is to make the system language independent. An Artist20 database is internationally accepted and acknowledged [37]. It comprises 20 songs of 20 singers of various genres. Since this database consists of only three female singers, we have considered only three male and female singers information for the task of gender classification. There are 100 clips for each male and female gender. The clips are with a sampling frequency of 44100 Hz. A sample data set of spectrograms are shown in Fig. 3. The length of each clip ranges from 2 to 5 seconds. The singers considered for the experiment have a different accent which makes the data-set very versatile and covers almost all the traits of the singers. This dataset is also divided into two parts for training 70% and testing 30%, respectively.

B. Results and Observations

The process of gender identification is quite easy with speech processing as pitch related features are sufficient to detect the same. Hence, the accuracy is around 97.40% with the suitable features that are extracted from the speech signal [38]. However, the above-mentioned accuracy is obtained for the studio-recorded speech where background noise is not considered. The system developed based on the traditional feature engineering approach gets failed if real-time noise and other speech gets involved. Hence, the modern popular convolutional neural networks have been used in the recent article and stated that an average of around 99% accuracy is obtained [39]. However, the speech considered for this work is recorded in environmentally controlled situations. It is very difficult to obtain that much level of accuracy in the case of gender identity for a singer. As the singer's vocals are always accompanied by instrumental sounds, the complexity of gender identification gets proportionately increased. Moreover, the singer intentionally changes his/her pitch since they trained their vocal cords accordingly. Based on this, one can say that the pitch related features alone could not perform well in the case of the singer's gender identification [14].

However, the supporting spectral and temporal features may support the pitch features to improve the performance of gender identification in the cast of singers. Hence, the Mel frequency cepstral coefficients (MFCCs) have been used as base-line supporting features as they have proven their capability in the modeling music signal. However, speech researchers know its importance as it has outperformed many speech-related tasks [40, 41]. Also, the first and second-order differentiations on MFCCs provide a new set of features called velocity and acceleration features of MFCCs. They are popularly called Δ and ΔΔMFCC features [42]. In many of the research works, it is mentioned that the Δ and ΔΔMFCCs carry temporal information of the signal. As the temporal information is much useful in discriminating the gender information, they have been added to support the spectral and pitch features [43].

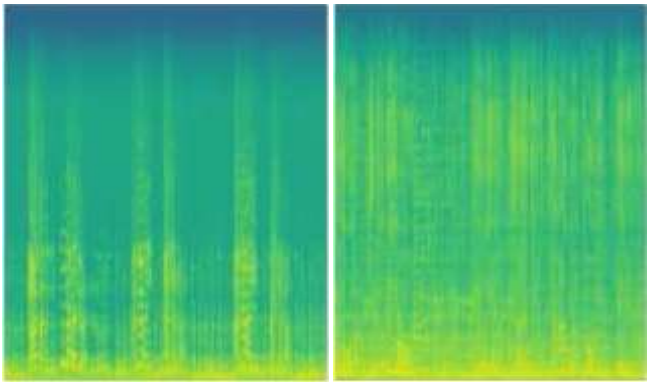


Fig. 3. A Sample Spectrogram Image that gives Some useful Discriminative Information of Male and Female Singers.

The consolidated features that are considered for this work are including MFCCs (13), pitch (4), ΔMFCCs (13), ΔΔMFCCs (13) forming a length of a 43-dimensional feature vector. The details of the features, acronyms, and length of each feature category are given in Table I. The second column represents the feature name, and the third one is the acronym that has been considered to represent the feature hereafter. The fourth column is the size of the respective feature. We have considered the features individually and different combinations that help estimate the relevant features for the task of the singer's gender identification.

The results obtained for the selective feature combinations are given in Table II and Fig. 4. The table gives complete information about the accuracies obtained for Indian and Western data clips. Moreover, three different classification models, such as ANN, RF, and CNN, have been considered to represent the data for both the categories. Initially, we experimented with the baseline MFCC features, obtaining an average accuracy of 63.60% and 69.42% for the Indian and Western clips, respectively. The percentage improvement with the combinations of {M+P}, {M+Δ+ΔΔ}, and {M+P+Δ+ΔΔ} is 16%, 3%, & 10% for Indian clips dataset, and 9%, 7%, & 6% for Western clips, respectively. However, the combination of {M+P+Δ+ΔΔ} is giving the best accuracy with the feature engineering process. The accuracies with the mentioned combination are 85.27% and 86.62% with Indian and Western clips, respectively. The values are average accuracies obtained for the ANN and RF. However, ANN and RF are found to be similar in their performances while classifying the genders for the above-mentioned feature combinations.

TABLE I. FEATURES, ACRONYMS, AND THEIR DIMENSIONAL SIZE THAT ARE CONSIDERED HEREAFTER

Sl. No.	Feature Name	Acronym	Size
1	MFCCs	M	13
2	Pitch	P	4
3	Δ MFCCs	Δ	13
4	Δ Δ MFCCs	Δ Δ	13

TABLE II. THE ACCURACY VALUES ARE OBTAINED USING THE VARIOUS COMBINATIONS OF FEATURE SETS FOR INDIAN AND WESTERN DATASETS

Features and Spectrograms	Accuracy (in %)					
	Indian			Western		
	ANN	*RF*	*CNN*	*ANN*	*RF*	*CNN*
M	62.34	64.86	-	69.29	69.54	-
M+P	74.82	73.24	-	76.34	75.25	-
M+Δ+ΔΔ	77.36	76.34	-	81.25	82.04	-
M+P+Δ+ΔΔ	83.72	86.82	-	85.19	88.05	-
GAFS<M+P+Δ+ΔΔ>	85.45	87.16	-	90.84	92.55	-
Spectrogram	-	-	***89.16***	-	-	***94.25***

There could be a chance of having some worthless feature dimensions in the selected combinational feature vector. Feature selection is one suitable approach to select the supporting feature dimensions and ignoring the rest, which may lead to an increase in the final accuracy. Hence, we applied a feature selection algorithm called genetic algorithm based feature selection (GAFS) to select the suitable features. The genetic algorithm comes under the category of evolutionary algorithms, which is purely based on randomness in the approach. The use of GAFS has given better performance on top of the best-combined feature vector. We have used the best combination {M+P+Δ+ΔΔ} to apply the GAFS algorithm mentioned in the 5th row of Table II, labeled GAFS<M+P+Δ+ΔΔ>. An increase of 1% and 5% have been obtained over the best-combined feature vector with the support of feature selection using GAFS.

Further, CNNs have been used to do experimentation to detect the accurate gender of the signer's voice. Spectrograms have been considered as they can discriminate the information related to male and female singers. Based on the visual differences observed in the spectrogram, an effort has been made to classify the gender with the support of CNN's. Table III gives detailed information about the hyperparameters and their values that are considered for the task of the singer's gender identification.

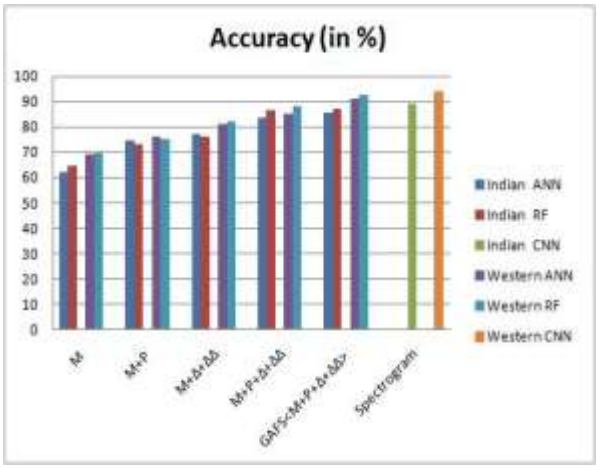


Fig. 4. The Graphical Representation of Accuracy Values that are Obtained using the Various Combinations of Feature sets for Indian and Western Datasets.

TABLE III. HYPERPARAMETERS ARE CONSIDERED FOR DESIGNING CNN FOR THE TASK OF THE SINGER'S GENDER CLASSIFICATION

Sl. No.	Parameter	Value
1	Batch size	8
2	#Channels	3 channels (RGB)
3	Filter size	3*3
4	Image size	128*128
5	#Convolution Layers	4
6	#Hidden layers	4
7	#Flatten layers	2
8	Softmax layer	1
9	#Output classes	2 (M & F)
10	Activation function(s)	ReLu
11	#Epochs	Around 250

A better accuracy has been obtained with the specified hyperparameters. CNN's outperform in many image processing tasks. Similarly, better performance has been observed in this case, as well. However, there is not much higher spike, which has been observed with the support of CNN's. A nominal improvement of 3% and 2.7% for the Indian and Western clips, respectively. However, CNN's can classify gender information with an accuracy of 89.16% and 94.25%, though there is complex background support by instrumentals. Hence, CNNs can be effectively utilized hereafter with various spectrogram models, which further could classify the singer's gender efficiently.

V. CONCLUSION

The process of the singer's gender identification will surely help the task of music information retrieval (MIR) and music recommender systems as well. The pitch alone features may not suffice to get better accuracy in the case of the singer's gender identification. An accuracy of 20% has been obtained using pitch features alone. The reason could be the support of complex background instrumentals involved in the case of music clips over speech signal processing. However, features that have been used in speech processing are effectively used in many music processing tasks as well. For instance, MFCCs are effectively utilized to model speech data and music also. Hence, spectral features MFCCs are suitable to give their support for pitch features to get better accuracy. Moreover, the temporal features obtained by applying first, and second-order differential equations on MFCCs resulting in velocity, and acceleration features. They are also useful to estimate the temporal variations in the signal effectively. It could be the reason for obtaining a considerable accuracy while using the combinational feature vector. It is also more important to omit worthless feature dimensions to avoid performance degradation. The support of recent CNNs is always effective in getting better accuracy over the traditional feature engineering process.

VI. FUTURE SCOPE

For future work, it is very important to establish a standard dataset for Indian singers. It may help in many of the MIR

tasks. The use of recurrent neural networks (RNNs) by feeding a one-dimensional signal could help in improving the accuracy as they outperform in many speech-related tasks. Hence, our future work focuses on the use of RNNs for gender identification. Moreover, it focuses on constructing an efficient dataset for Indian audio clips. Since the structure of Indian songs is completely different from Western clips, it is highly essential to construct the same. Further, we may focus on the task of singer identification using gender classification as a fundamental step. It means the task of singer identification can be effectively done if a two-level classification model is proposed.

AUTHORSHIP CONTRIBUTION STATEMENT

Mukkamala S N V Jitendra: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft & editing. Dr. Y. Radhika: Supervision, Conceptualization, Writing - review & editing.

ACKNOWLEDGMENT

We acknowledge support from the Department of Computer and Engineering, Gandhi Institute of Technology and Management (GITAM) Deemed to be University, Vishakhapatnam for guidance, reviews, valuable suggestions, and very useful discussions for all the support being extended to carry out this research work.

REFERENCES

[1] Murthy, YV Srinivasa, and Shashidhar G. Koolagudi. "Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review." ACM Computing Surveys (CSUR), vol. 51(3), 2018, pp.1-46.

[2] Cai, Wei, Qiang Li, and Xin Guan. "Automatic singer identification based on auditory features." In 2011 seventh international conference on natural computation, vol. 3, IEEE 2011, pp. 1624-1628.

[3] Ter Bogt, Tom FM, Juul Mulder, Quinten AW Raaijmakers, and Saoirse Nic Gabhainn. "Moved by music: A typology of music listeners." Psychology of Music, vol.39(2), 2011, pp. 147-163.

[4] Downie, J. Stephen. "Music information retrieval." Annual review of information science and technology, vol.37 (1), 2003, pp. 295-340.

[5] Schedl, Markus, Peter Knees, and Gerhard Widmer. "Investigating web-based approaches to revealing prototypical music artists in genre taxonomies." In 2006 1st International Conference on Digital Information Management, IEEE 2006, pp. 519-524.

[6] Harb, Hadi, and Liming Chen. "Voice-based gender identification in multimedia applications." Journal of intelligent information systems, vol. 24(2), 2005, pp. 179-198.

[7] Alsharhan, E. and Ramsay, A., "Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions". Information Processing & Management, vol. 56(2), 2019, pp.343-353.

[8] Nakano, Tomoyasu, Kazuyoshi Yoshii, and Masataka Goto. "Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity." In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 5202-5206.

[9] Hu, Yakun, Dapeng Wu, and Antonio Nucci. "Pitch-based gender identification with two-stage classification." Security and Communication Networks, vol. 5(2), 2012, pp.211-225.

[10] Qian, Kun, Zixing Zhang, Fabien Ringeval, and Björn Schuller. "Bird sounds classification by large scale acoustic features and extreme learning machine." In 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2015, pp. 1317-1321.

[11] Huss, "Vocal Pitch Range and Habitual Pitch Level: The Study of Normal College Age Speakers" (1983). Master's Theses. 1590.

[12] Titze, I. R., & Martin, D. W. ,"Principles of voice production". The Journal of the Acoustical Society of America, vol. 104, 1998, pp. 1148.

[13] Barkana, Buket D., and Jingcheng Zhou. "A new pitch-range based feature set for a speaker's age and gender classification." Applied Acoustics, vol. 98, 2015, pp.52-61.

[14] Weninger, F., Wöllmer, M., & Schuller, B, "Automatic assessment of singer traits in popular music: Gender, age, height and race". Paper presented at the Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, pp. 37-42.

[15] Rao, K. Sreenivasa, and Shashidhar G. Koolagudi. "Identification of Hindi dialects and emotions using spectral and prosodic features of speech." IJSCI: International Journal of Systemics, Cybernetics and Informatics, vol. 9(4), 2011, pp. 24-33.

[16] Gaikwad, Santosh, Bharti Gawali, and S. C. Mehrotra. "Gender identification using SVM with combination of MFCC." Advances in Computational Research, vol. 4(1), 2012, pp. 69-73.

[17] Zeng, Yu-Min, Zhen-Yang Wu, Tiago Falk, and Wai-Yip Chan. "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech." In 2006 International Conference on Machine Learning and Cybernetics, IEEE, 2006, pp. 3376-3379.

[18] Chen, Gang, Xue Feng, Yen-Liang Shue, and Abeer Alwan. "On using voice source measures in automatic gender classification of children's speech." In Eleventh Annual Conference of the International Speech Communication Association, 2010, pp. 673-676.

[19] Sedaghi, M. "A comparative study of gender and age classification in speech signals". Iranian Journal of Electrical and Electronic Engineering, vol. 5(1), 2009, pp. 1-12.

[20] Alsulaiman, Mansour, Zulfiqar Ali, and Ghulam Muhammad. "Gender classification with voice intensity." In 2011 UKSim 5th European Symposium on Computer Modeling and Simulation, IEEE, 2011, pp. 205-209.

[21] Alsulaiman, Mansour, Zulfiqar Ali, and Ghulam Muhammad. "Voice intensity based gender classification by using Simpson's rule with SVM." In 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, 2012, pp. 552-555.

[22] Murthy, YV Srinivasa, and Shashidhar G. Koolagudi. "Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS)." Expert Systems with Applications, vol. 106, 2018, pp. 77-91.

[23] Murthy, YV Srinivasa, and Shashidhar G. Koolagudi. "Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations." In 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2015, pp. 1271-1276.

[24] Murthy, Y. VS, T. K. R. Jeshventh, M. Zoeb, M. Saumyadip, and G. K. Shashidhar. "Singer identification from smaller snippets of audio clips using acoustic features and DNNs." In 2018 Eleventh International Conference on Contemporary Computing (IC3), . IEEE, 2018, pp. 1-6.

[25] Sharma, Rahul, YV Srinivasa Murthy, and Shashidhar G. Koolagudi. "Audio songs classification based on music patterns." In Proceedings of the second international conference on computer and communication technologies, Springer, New Delhi, 2016, vol. 381, pp. 157-166.

[26] Thomas, Matthew, YV Srinivasa Murthy, and Shashidhar G. Koolagudi. "Detection of largest possible repeated patterns in indian audio songs using spectral features." In 2016 IEEE Canadian conference on electrical and computer engineering (CCECE), IEEE, 2016. pp. 1-5.

[27] Jitendra, M. S. N. V., & Radhika, Y. "A review: Music feature extraction from an audio signal". International Journal of Advanced Trends in Computer Science and Engineering, vol. 9(2), 2020, pp. 973-980.

[28] Mason, J. S., and X. Zhang. "Velocity and acceleration features in speaker recognition." In Acoustics, Speech, and Signal Processing, IEEE International Conference on, pp. 3673-3674. IEEE Computer Society, 1991.

[29] Furui, Sadaoki. "Comparison of speaker recognition methods using statistical features and dynamic features." IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29(3), pp. 342-350, 1981.

[30] Qawaqneh, Zakariya, Arafat Abu Mallouh, and Buket D. Barkana. "Deep neural network framework and transformed MFCCs for speaker's

age and gender classification." Knowledge-Based Systems, vol. 115, pp. 5-14, 2017.

[31] Biswas, Roshni, YV Srinivasa Murthy, Shashidhar G. Koolagudi, and Swaroop G. Vishnu. "Objective Assessment of Pitch Accuracy in Equal-Tempered Vocal Music Using Signal Processing Approaches." In Smart Computing Paradigms: New Progresses and Challenges, vol. 766, pp. 161-168. Springer, Singapore, 2020.

[32] Murthy, YV Srinivasa, Shashidhar G. Koolagudi, and Vishnu G. Swaroop. "Vocal and Non-vocal Segmentation based on the Analysis of Formant Structure." In 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1-6. IEEE, 2017.

[33] Zeng, Yuni, Hua Mao, Dezhong Peng, and Zhang Yi. "Spectrogram based multi-task audio classification." Multimedia Tools and Applications, vol. 78(3), pp. 3705-3722, 2019.

[34] Russo, Mladen, Luka Kraljević, Maja Stella, and Marjan Sikora. "Cochleogram-based approach for detecting perceived emotions in music." Information Processing & Management, vol. 57(5), pp.102270. 2020.

[35] Costa, Yandre MG, Luiz S. Oliveira, and Carlos N. Silla Jr. "An evaluation of convolutional neural networks for music classification using spectrograms." Applied soft computing, vol. 52, pp. 28-38, 2017.

[36] Badshah, Abdul Malik, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. "Speech emotion recognition from spectrograms with deep convolutional neural network." In 2017 international conference on platform technology and service (PlatCon), pp. 1-5. IEEE, 2017.

[37] Ellis, Daniel PW. "Classifying music audio with timbral and chroma features". Paper presented at the Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, pp. 339-340.

[38] Grimaldi, Marco, and Fred Cummins. "Speaker identification using instantaneous frequencies." IEEE transactions on audio, speech, and language processing, vol. 16(6), pp. 1097-1111, 2008.

[39] Rami S. Alkhawaldeh, "DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network", Scientific Programming, vol. 2019, Article ID 7213717, pp.1-12, 2019. https://doi.org/10.1155/2019/721371.

[40] Alam, Md Jahangir, Tomi Kinnunen, Patrick Kenny, Pierre Ouellet, and Douglas O'Shaughnessy. "Multitaper MFCC and PLP features for speaker verification using i-vectors." Speech communication, vol. 55(2), pp. 237-251, 2013.

[41] Gupta, Shruti, Md Shah Fahad, and Akshay Deepak. "Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition." Multimedia Tools and Applications, vol. 79, pp. 23347-23365, 2020.

[42] Hossan, Md Afzal, Sheeraz Memon, and Mark A. Gregory. "A novel approach for MFCC feature extraction." In 2010 4th International Conference on Signal Processing and Communication Systems, pp. 1-5. IEEE, 2010.

[43] Hu, Maodi, Yunhong Wang, Zhaoxiang Zhang, and Yiding Wang. "Combining spatial and temporal information for gait based gender classification." In 2010 20th International Conference on Pattern Recognition, pp. 3679-3682. IEEE, 2010.

[44] Kaluri, Rajesh, and P. Reddy. "Sign gesture recognition using modified region growing algorithm and adaptive genetic fuzzy classifier." Int J Intell Eng Syst 9 (2016): pp. 225-233.

[45] Kaluri, Rajesh, and C. H. Pradeep. "An enhanced framework for sign gesture recognition using hidden Markov model and adaptive histogram technique." Int J Intell Eng Syst 10 (2017):pp. 11-19.

[46] Kaluri, Rajesh, and Pradeep Reddy CH. "Optimized feature extraction for precise sign gesture recognition using self-improved genetic algorithm." Int. J. Eng. Technol. Innov 8, no. 1 (2018):pp. 25-37.

[47] Reddy, G. Thippa, M. Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, and Thar Baker. "Analysis of dimensionality reduction techniques on big data." IEEE Access 8 (2020): pp. 54776-54788.

An Hybrid Approach for Cost Effective Prediction of Software Defects

Satya Srinivas Maddipati[1]
Research Scholar, CSE Department
Koneru Lakshmaiah Education Foundations
Guntur, India

Malladi Srinivas[2]
Professor, CSE Department
Koneru Lakshmaiah Education Foundations
Guntur, India

***Abstract*—Identifying software defects during early stages of Software Development life cycle reduces the project effort and cost. Hence there is a lot of research done in finding defective proneness of a software module using machine learning approaches. The main problems with software defect data are cost effective and imbalance. Cost effective problem refers to predicting defective module as non defective induces high penalty compared to predicting non defective module as defective. In our work, we are proposing a hybrid approach to address cost effective problem in Software defect data. To address cost effective problem, we used bagging technique with Artificial Neuro Fuzzy Inference system as base classifier. In addition to that, we also addressed Class Imbalance & High dimensionality problems using Artificial Neuro Fuzzy inference system & principle component analysis respectively. We conducted experiments on software defect datasets, downloaded from NASA dataset repository using our proposed approach and compared with approaches mentioned in literature survey. We observed Area under ROC curve (AuC) for proposed approach was improved approximately 15% compared with highly efficient approach mentioned in literature survey.**

Keywords*—Cost effective problem; principle component analysis; adaptive neuro fuzzy inference system; area under ROC curve

I. Introduction

Software Development process involves Requirement specification, Design, Implementation and Testing. During each phase of software development, reviews will be conducted to assess the progress and quality of software. The quality of software depends on defects found in the software. Defect is a condition that doesn't meet user requirement, specified in requirement specification. If a defect is found during late stages of software development i.e. during software maintenance, the penalty is very high. To reduce this penalty, the defective proneness must be identified in advance [27].

According to Boehm, the cost of fixing errors increased gradually as the software development progress. If we consider cost of fixing error during requirement phase as 1 unit, then the cost of fixing error in design phase will be 3-8 units, implementation phase will be 7 to 16 units, integration & testing phase will be 21 to 78 units and maintenance phase will be 29 to more than 1500 units. This motivates application of machine learning techniques in early stage identification of software defects [28]. Fig. 1 shows soft escalation of defect resolving during various phases of software development life cycle.

A. Machine Learning Techniques

Various Machine learning techniques such as K nearest neighbours, Support Vector machines, Decision Trees, Bayesian Networks and etc. are used to identify software defects.

B. Approaches for Software Defect Prediction (SDP)

1) Decision trees: Decision Trees are used as early classifier techniques for software defects. In a decision tree, the attribute with less impurity value is selected as root node. There are three measures for impurity 1) Entropy 2) Gini Index 3) Misclassification error. Decision tree will output whether the module is defective prone or not, based on input attributes like IO Comments, Cyclometric complexities etc. Fig. 2 shows Decision Tree constructed on cm1 dataset.

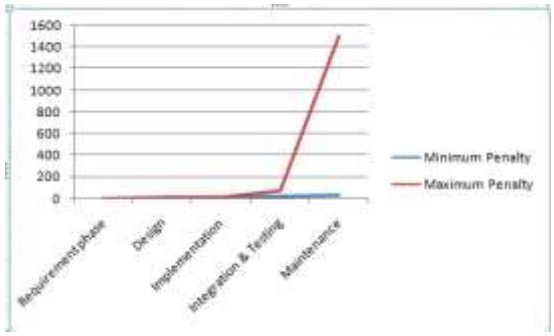


Fig. 1. Cost Escalation for Defect Solving during Phase of Software Development.

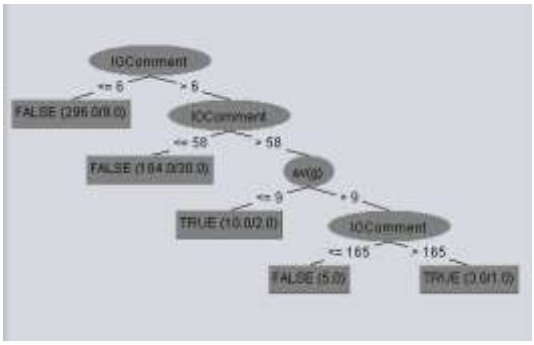


Fig. 2. Decision Tree on CM1 Dataset.

2) Bayesian Classifiers: Bayesian classifier uses Baye's theorem to classify unknown sample. It comes under lazy classifier. According to Bayes theorem, Conditional probability $P(Y=y_i/X=x_i)$ is defined as:

$$P(Y = yi/X = xi) = \frac{P(X = xi/Y = yi) * P(Y = yi)}{P(X = xi)}$$

There are two types of Bayesian classifiers: 1) Naive Baye's classifier 2) Bayesian Belief Networks.

Naive Baye's classifier: In Naive Baye's classifier, the given unknown sample is considered as 'X'. The classifier finds the posterior probability P (Defect=Yes/X) and P(Defect=No/X) for given sample 'X'.

If P(Defect=Yes/X) > P(Defect=No/X), then classifier outputs the sample 'X' as defective. Otherwise it outputs the given sample 'X' as non defective. The drawback with Naive Baye's classifier is, it assumes the target variable (Defect) is independent on input variables.

Bayesian Belief Networks: In Bayesian Belief Networks, There are two components: 1) Direct Acyclic Graph (DAG); 2) Probability table DAG encodes the relationship between attributes into a graph. Probability table comprises of posterior probabilities dependent on their parent attributes. Fig. 3 represents the DAG, constructed on cm1 dataset.

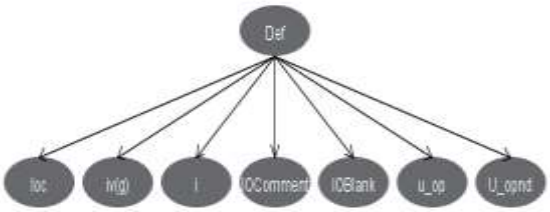


Fig. 3. Directed Acyclic Graph for Software Defect Prediction.

3) Support vector machines: Support vector machines are one of the popular classifier technique for regression and classification problems. Binary Support Vector Machines solves classification tasks while support vector regression solves regression tasks. Software defect prediction is a classification problem and hence Binary Support Vector Machines are used to classify the module as defective or non-defective. In support vector machines, there exist a boundary function that classify sample. There are various types of boundary functions like Linear, Polynomial, Radial basis, ANOVA and etc.

Polynomial Support Vector Machine:

Defect = -0.0062 * (normalized) loc+ 0.0043 * (normalized) iv(g) +0.0044 * (normalized) i+0.012 * (normalized) IOComment + -0.0021 * (normalized) IOBlank + 0.0004 * (normalized) u_op +-0.0054 * (normalized) U_opnd - 1.0005.

Multi Layer Perceptrons: Multi layer perceptrons are the neural networks which comprises Processing units, called neurons, organized in multiple layers. These neurons are having computing capabilities on inputs, receiving from previous layers, and propagate output to the next layers. These neurons are connected by weighted edges. Each neurons applies activation functions on the inputs along with threshold and produces output signals. There are various activation functions like Threshold, sigmoid, Tangible and etc.

Table I illustrates the architecture of multi layer perceptrons along with nodes, connections and their weights.

Artificial Neuro Fuzzy Inference System:

Artificial Neuro Fuzzy Inference System: ANFIS is a five layered architecture used for classification tasks. Satya srinivas et al. [26] proposed Artificial Neuro Fuzzy Inference System for Software defect prediction.

ANFIS generates Sugeno Fuzzy Inference system as output for classification task. In ANFIS input attributes are fuzzified and target attribute is defuzzified. Intially subtractive clustering method is used to generate Sugeno fuzzy inference system. The premise and consequent parameters in Sugeno Fuzzy inference system are trained used training data. Here training rate parameter must be set to appropriate value. Setting high training rate parameter converges the ANFIS model into unstable state. Setting low training rate parameter creates high complexity model.

4) Cost effective learning: Misclassifying some class samples results high penalty compared to misclassification of other classes. For example, in software Defect prediction, misclassifying defective module as non defective imposes high penalty compared to misclassifying non defective module as defective. If a defect was found during later stages of software development, it imposes high penalty and hence pronable defective module should not be misclassified as non defective even though non defective module was misclassified as defective. This error cost escalation was shown in Fig. 1.

5) Ensemble learning: Ensemble learning is the process of constructing multiple classifiers and combining them to improve the accuracy for classification problems. Some of the ensembling techniques are Simple voting, Average voting, Bagging, Boosting and etc. In simple voting, each classifier will vote for an output value. The output, value with high number of votes, considered as actual output. In average voting, the average value of output of each classifier is considered as actual output. This technique is suitable for regression tasks. In Bagging, the dataset is sampled into equal size subsets of data and a classifier is constructed with each subset. Finally each classifier will vote for output value. Bagging and Boosting techniques improves the performance of classifiers by constructing multiple classifiers.

In Bagging, classifiers are constructed in sequence. The samples which are incorrectly classified are given with higher weight for construction of next classifier. This procedure is repeated until required accuracy obtained or maximum numbers of classifiers were constructed.

TABLE I. MULTI LAYER PERCEPTRONS NODES, CONNECTIONS AND WEIGHTS

Node	Inputs	Weights	Node	Inputs	Weights	Node	Inputs	Weights
0	Threshold	-3.79506	2	Attrib u_op	4.04982	4	Attrib u_op	3.31857
0	Node 2	4.53581	2	Attrib U_opnd	4.93275	4	Attrib U_opnd	4.17034
0	Node 3	3.66528	3	Threshold	-4.9707	5	Threshold	-9.06268
0	Node 4	5.20758	3	Attrib loc	2.20562	5	Attrib loc	-3.81299
0	Node 5	3.18853	3	Attrib iv(g)	-7.07055	5	Attrib iv(g)	-1.35707
1	Threshold	3.79522	3	Attrib i	-6.27414	5	Attrib i	-0.84483
1	Node 2	-4.50871	3	Attrib IOComment	-2.75564	5	Attrib IOComment	-4.81926
1	Node 3	-3.66711	3	Attrib IOBlank	-1.19511	5	Attrib IOBlank	3.59625
1	Node 4	-5.23767	3	Attrib u_op	0.39515	5	Attrib u_op	-5.56998
1	Node 5	-3.18676	3	Attrib U_opnd	3.76011	5	Attrib U_opnd	-1.61555
2	Threshold	-2.83258	4	Threshold	-2.79212		Class FALSE	
2	Attrib loc	4.66117	4	Attrib loc	4.39579		Input	
2	Attrib iv(g)	0.10108	4	Attrib iv(g)	2.53554		Node 0	
2	Attrib i	-1.00049	4	Attrib i	-0.01669		Class TRUE	
2	Attrib IOComment	-8.27215	4	Attrib IOComment	-9.35516		Input	
2	Attrib IOBlank	0.286789	4	Attrib IOBlank	0.769326		Node 1	

In this paper, we are applying hybrid approach to overcome cost effective problem in SDP. Section II presents literature survey on SDP. In Section III, we designed methodology using hybrid approach for SDP. Section IV Presents the results by applying proposed methodology on SDP.

II. LITERATURE SURVEY

Yan Naung Soe et al. proposed Random Forest algorithm on Software Defect Prediction and compared the performance of Random forest algorithm with other machine learning techniques. They concluded that maximum accuracy is 99.59 and minimum accuracy is 85.96[1]. Taek Lee et al. proposed micro interaction metrics, such as browsing events, file editing, for prediction of software defects and observed high accuracy by combining these metrics with existing metrics in cost effective manner [2]. Fei Wu et al. proposed a cost-sensitive local collaborative representation (CLCR) approach for software defect prediction and concluded that accuracy has been increased with proposed approach [3]. Jinsheng Ren et.al proposed asymmetric kernel principle component analysis for solving class imbalance problem in software defect prediction. They evaluated the validity of their proposed model using *F*-measure, Friedman's test, and Tukey's test [4].

Ayse Tosun et al. proposed decision threshold optimization on Naive Bayes classifier to find best threshold that separate defective and non-defective samples in software defect data [5]. Ming Cheng et al. proposed semi supervised approach for identification of software defects. Their proposed model evaluates the confidence probability of unlabelled sample to predict class labels. They considered different misclassification cost to improve classifier performance [6]. Igor Ibarguren et al proposed consolidated tree construction that ensembles weights of misclassification in training of classifier. They showed that consolidated tree construction performs better than other rule based classifiers [7]. Yuanxun Shao et.al proposed weighted associative classification for addressing imbalance problem in software defect prediction. They determined weights of features using correlation analysis. They proved GMean measure has been increased with their approach [8]. Shuo Feng et al. proposed complexity based over sampling technique to address data imbalance problem in identification of software defects [9]. Rakesh Rana et al. proposed Bayesian Inference method for software defect prediction to analyse inflow distribution of defects. This technique has been used for early detection of software defects in large software projects [10].

Guisheng Fan et.al proposed attention based recurrent neural networks for software defect prediction. Their experimental results shows that the proposed model increases F1 score by 14% and AUC by 7% [11]. Sushant Kumar et al. proposed Deep representation and ensemble learning for Software defect prediction. They conducted experiments on 12 NASA Dataset repositories. Among 12 datasets, F Measure has been increased for 8 datasets, ROC values has been increased for 6 datasets, PRC values has been increased for 12 datasets and MCC values has been increased for 11 datasets[12]. Rodrigo et al. proposed ensemble of clustering using Particle Swarm Optimization for prediction of Software defects and concluded that prediction quality has been increased [13]. Shamsul Huda et al. proposed ensemble over sampling algorithm for prediction of software defects [14]. Shanthini. et al. proposed Ensemble SVM approach for prediction of software defects[15]. Nageswara Rao et.al proposed Ensemble Bayesian networks for prediction of Software Defects and proved that their proposed model have high true positive rate compared to traditional methods [16]. Steven Young et al. proposed deep super learner for Just in time defect prediction. They used bagging of random forests

and concluded that F1 score was improved for 5 of 6 projects [17]. Arvinder Kaur et al investigated different ensemble techniques such as Boosting, Bagging and Rotation forest in prediction of software defects. They conducted Wilcoxon signed rank test to prove ensemble techniques outperforms traditional techniques in generalization of results [18]. Thanh Tung et al proposed ensemble model by combining sampling technique with common classification technique to improve the performance of classifier [19].

Jaroslaw Hryszko et al investigated the effect of Software defect in modules on Quality assurance of Software. Their investigation proved that quality assurance cost can be reduced by 30% with their proposed approach [20]. Kazuya Tanaka et al focused on usage of auto-sklearn tool that automatically selects appropriate prediction model for data pre processing and classification in software defect prediction. This tool presents random forest is the best model in various machine learning techniques [21]. Pradeep Singh proposed stacking based framework, in which he combined class balancing technique SMOTE with ensemble classifiers to predict software defects. He concluded that the accuracy of stacking based model increased compared to traditional approaches used in their literature survey [22]. Haitao He et al proposed Ensemble RIPPER classifier for software defect prediction. In their research, they applied Principle component analysis for dimensionality reduction, Adaptive Synthetic sampling for balancing the dataset and RIPPER model for classification. They concluded that classification error has been reduced with their proposed model [23]. Zhiqiang Li et al proposed ensemble multiple kernel correlation alignment for heterogeneous defect prediction and they concluded ensemble approach outperforms remaining competing methods [24]. Xin Xia et al proposed Hybrid model reconstruction (HYDRA) approach for Software defect prediction. It consists of two phase's Genetic algorithm followed by Ensemble learning. They concluded that HYDRA improves F1 score of Zero-R base classifier [25].

In prediction of software defects, some researchers addressed class imbalance problem and someone addressed high dimensionality problem. But In this research work, we are addressing cost effective problem in SDP.

III. METHODOLOGY

In this paper, we are proposing Ensemble approach of Adaptive Neuro Fuzzy Inference system for prediction of Software defects for cost effective learning. In step 1, we are performing Synthetic Minority oversampling technique (SMOTE) to balance the dataset. In step 2, Dimensionality reduction will be performed to reduce the dataset. Here, we are proposing Principle component analysis (PCA) for dimensionality reduction. In step 3, multiple ANFIS classifiers will be constructed for ensemble approach. In step 4, Aggregation will be performed on votes given by multiple ANFIS classifiers and it produces the actual output.

In our research work, we considered data from NASA dataset repository. The dataset is neither noisy nor in complete but imbalanced. To remove imbalance, we are applying SMOTE technique and to overcome for high dimensionality problem we are applying PCA. Fig. 4 represents the proposed methodology for SDP.

Algorithm:

Step 1: Apply Synthetic Minority Over Sampling Technique for Class Balance.

- 1.1 Choose a random sample from minority class.
- 1.2 Identify k-nearest neighbours from chosen sample
- 1.3 For each neighbour sample
  - 1.3.1 construct a line from chosen sample to nearest neighbour
  - 1.3.2 Add more number samples by picking of points on the line
- 1.4 Repeat steps 1.1 to 1.3 until two classes samples are equal.

Step 2: Apply Dimensionality reduction using PCA

- 2.1 Perform Z score normalization on data.
  Z-score = $(x_i - \mu)/\sigma$
- 2.2 Create a covariance matrix for eigen decomposition.
- 2.3 select principle components with high relevance.

Step 3: Construct classifier using Artificial Neuro Fuzzy Inference system

- 3.1 Fuzzify input variable
- 3.2 Apply membership function on input variable
- 3.3 Calculate weighted average
- 3.4 calculate contribution of each fuzzy rule
- 3.5 Output sum of all incoming signals.

Step 4: Repeat steps 1 to Step 3 for multiple times (Possibly odd number of times).

Step 5: find number of votes for each class from multiple classifiers

Step 6: Output the class variable based on number of votes(High number of votes)

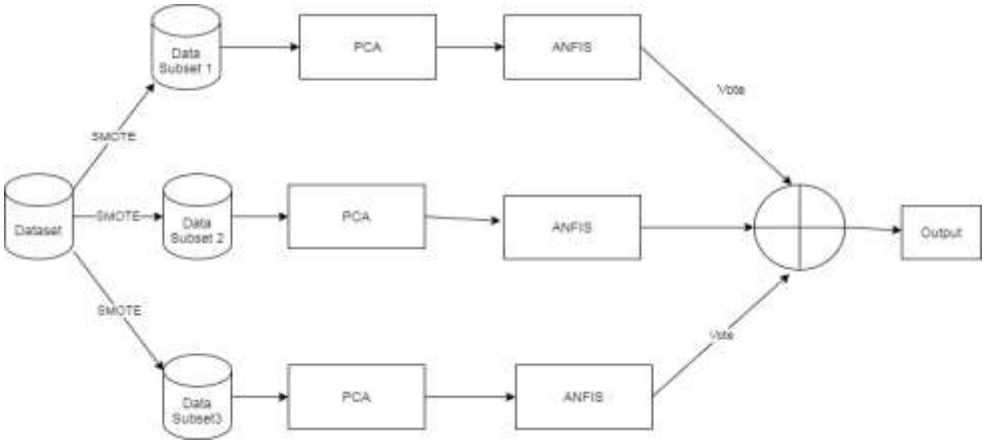


Fig. 4. Proposed Methodology for Software Defect Prediction.

IV. Experimentation and Results

In this work, we addressed class imbalance problem using Synthetic Minority Oversampling Technique. After balancing the data, we applied principle component analysis for dimensionality reduction. In Table II, we are projecting few principle component values constructed on cm1 dataset, downloaded from NASA dataset repository.

The reduced dataset is used to construct classifier using Adaptive Neuro Fuzzy Inference System. In this work, we applied AdaBoost Ensemble learning technique with ANFIS as base classifier. The performance of a classifier, constructed from imbalance data, can be measured using AuC (Area under ROC Curves). Receiver Operating Characteristics curves are constructed by plotting True positive rate against False positive rate. Area under this ROC curve is considered as a performance metric in our research work.

We applied cost sensitive approach to our classifier. In cost sensitive approach, the cost values are derived from imbalance nature of data. We found cost sensitive approach improves the performance of classifier. We applied our proposed model on various software defect datasets cm1, pc1, kc1 and jm1. In the Table III, We are comparing the AuC values of our proposed model with results of methods proposed in literature survey. Fig. 5 to 16 compares the ROC curves of various techniques discussed in literature survey with proposed methodology.

TABLE II. Principle Component Values on Software Defect Prediction

Attribute	PC1	PC2	PC3	PC4	PC5
loc	0.246323	0.031239	-0.08786	0.161044	-0.04483
V.g	0.243657	0.057173	0.103265	0.084628	-0.0588
ev.g.	0.205853	0.053904	0.118336	0.050775	-0.11766
iv.g.	0.233961	0.073498	0.029074	0.145675	0.010602
n	0.250283	-0.00128	-0.06655	-0.00328	0.006294
v	0.250585	0.043622	-0.04154	0.067938	0.037045
l	-0.09998	0.657384	-0.0066	0.191268	0.294755
d	0.211336	-0.14493	0.234525	-0.34617	-0.30322
i	0.193738	-0.11365	-0.60919	0.121661	0.10463
e	0.219659	0.167325	0.392074	0.132239	-0.0035
b	0.247381	0.132791	-0.07955	-0.00411	-0.01886
t	0.219659	0.167332	0.392072	0.132234	-0.0035
IOCode	0.199326	0.034844	0.132263	-0.30874	0.501051
IOComment	0.209861	0.048167	-0.12006	0.257712	-0.14999
IOBlank	0.188142	-0.05663	-0.02223	-0.46712	0.52168
locCC	-0.01344	0.623327	-0.27367	-0.513	-0.40025
u_op	0.218739	-0.21843	0.018721	-0.27123	-0.27324
U_opnd	0.238652	-0.03448	-0.31957	0.088945	0.053283
total_op	0.249928	-0.00192	-0.06993	0.008577	-0.01499
total_opnd	0.248372	0.000253	-0.06084	-0.0222	0.039162
branchCount	0.243334	0.040645	0.048155	0.08437	-0.06212

TABLE III. AUC VALUES OF VARIOUS MODELS WITH AND WITHOUT COST SENSITIVE TECHNIQUES

Dataset	CM1		PC1		KC1		JM1	
Base Classifier	**without Cost Sentive**	**With Cost sensitive**	**without Cost Sentive**	**With Cost sensitive**	**without Cost Sentive**	**With Cost sensitive**	**without Cost Sentive**	**With Cost sensitive**
J-48	0.53	0.56	0.49	0.53	0.48	0.54	0.49	0.55
Random Forest	0.53	0.74	0.51	0.71	0.5	0.72	0.51	0.69
SVM	0.49	0.55	0.48	0.53	0.5	0.53	0.51	0.57
K-NN	0.51	0.54	0.5	0.55	0.49	0.56	0.51	0.55
MLP	0.5	0.55	0.49	0.56	0.47	0.58	0.49	0.56
ANFIS	0.69	0.74	0.68	0.73	0.69	0.75	0.71	0.75

ROC Curves

J-48 (Decision Tree)

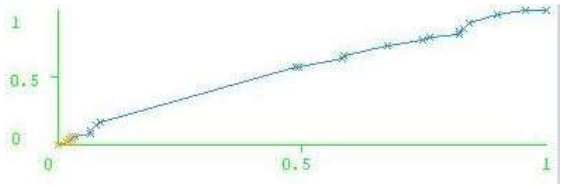


Fig. 5. ROC Curve using J-48 without Cost Sensitive.

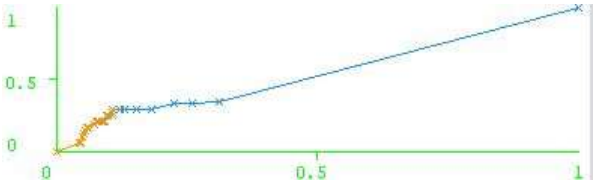


Fig. 6. ROC Curve using J-48 with Cost Sensitive.

Random Forest (RF)

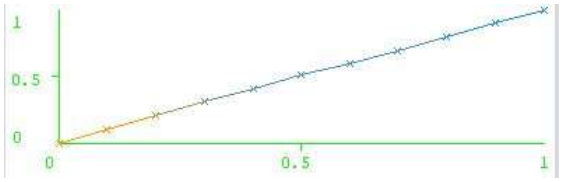


Fig. 7. ROC Curve using RF without Cost Sensitive.

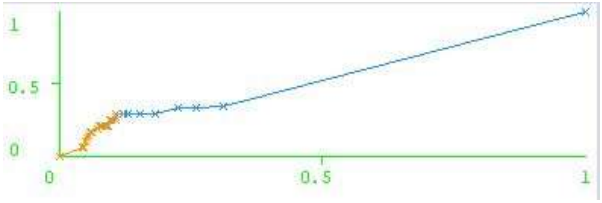


Fig. 8. ROC Curve using RF with Cost Sensitive.

Support Vector Machines (SVM)

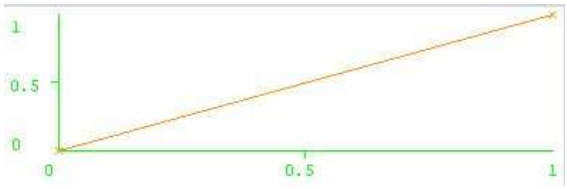


Fig. 9. ROC Curve using SVM without Cost Sensitive.

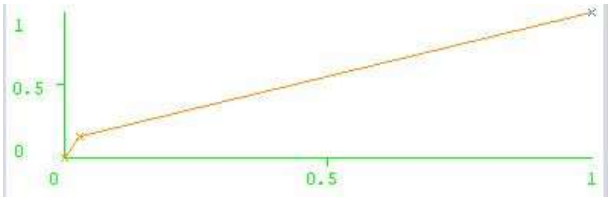


Fig. 10. ROC Curve using SVM with Cost Sensitive.

K Nearest Neighbour (KNN)

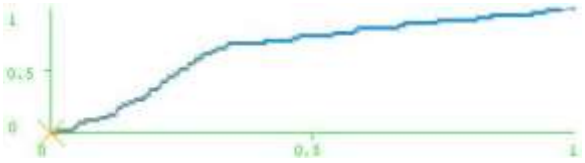


Fig. 11. ROC Curve using KNN without Cost Sensitive.

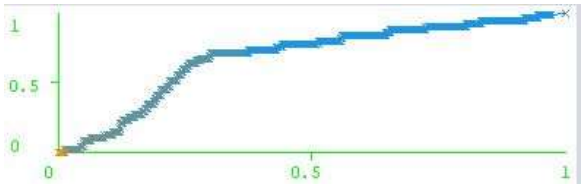


Fig. 12. ROC Curve using KNN with Cost Sensitive.

Multi Layer Perceptrons (MLP)

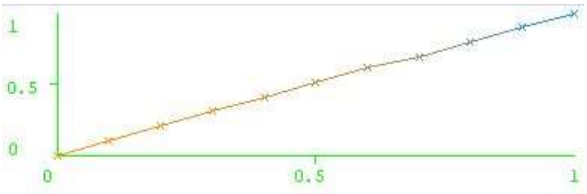


Fig. 13. ROC Curve using MLP without Cost Sensitive.

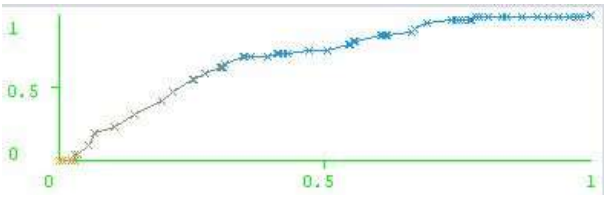


Fig. 14. ROC Curve using MLP with Cost Sensitive.

Adaptive Neuro Fuzzy Inference System (ANFIS)

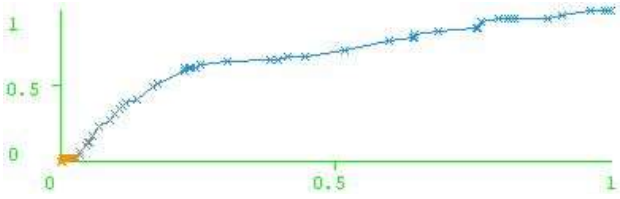


Fig. 15. ROC Curve using ANFIS without Cost Sensitive.

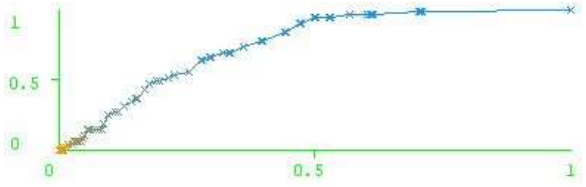


Fig. 16. ROC Curve using ANFIS with Cost Sensitive.

V. Conclusion and Future Work

In this work, we proposed an hybrid approach cost effective problems in software defect prediction. To reduce number of dimensions, we applied Principle Component Analysis. Ensemble ANFIS were constructed for cost effective learning of software defects. We compared the performance of proposed models, with algorithms specified in literature survey, using AuC values. Our proposed model got approximately 15% high Auc values over all datasets. As a future work, we can improve the AuC values by addressing High dimensionality, Class Imbalance, Cost effective Problems in SDP.

References

[1] Yan Naung Soe , Paulus Insap Santosa, Rudy Hartanto, "Software Defect Prediction Using Random Forest Algorithm", 2018 12th South East Asian Technical University Consortium (SEATUC), DOI: 10.1109/SEATUC.2018.8788881.

[2] Taek Lee, Jaechang Nam, Donggyun Han, Sunghun Kim, Hoh Peter In," Developer Micro Interaction Metrics for Software Defect Prediction", IEEE Transactions on Software Engineering,42(11).

[3] Fei Wu; Xiao-Yuan Jing; Xiwei Dong; Jicheng Cao; Baowen Xu; Shi Ying "Cost-Sensitive Local Collaborative Representation for Software Defect Prediction", 2016 International Conference on Software Analysis, Testing and Evolution (SATE), DOI: 10.1109/SATE.2016.24.

[4] Jinsheng Ren, Ke Qin, Ying Ma, Guangchun Luo," On Software Defect Prediction Using Machine Learning",Journal of Applied Mathematics,2014.

[5] Ayse Tosun, Ayse Bener, "Reducing false alarms in software defect prediction by decision threshold optimization", 2009 3rd International Symposium on Empirical Software Engineering and Measurement. DOI: 10.1109/ESEM.2009.5316006.

[6] CHENG Ming, WU Guoqing, YUAN Mengting and WAN Hongyan, "Semi-supervised Software Defect Prediction Using Task-Driven Dictionary Learning", Chinese Journal of Electronics,25(6).

[7] Igor Ibarguren, Jesus M.P´erez ´, Javier Mugerza , Daniel Rodriguez, Rachel Harrison" , The Consolidated Tree Construction Algorithm in Imbalanced Defect Prediction Datasets", 2017 IEEE Congress on Evolutionary Computation (CEC), DOI: 10.1109/CEC.2017.7969629.

[8] Yuanxun Shao,Bin Liu,Shihai Wang,Guoqi Li," Software defect prediction based on correlation weighted class association rule mining", Knowledge-Based Systems,196.

[9] Shuo Feng, Jacky Keung, Xiao Yu, Yan Xiao, Kwabena Ebo Bennin, Md Alamgir Kabir, Miao Zhang, "COSTE: Complexity-based OverSampling TEchnique to alleviate the class imbalance problem in software defect prediction", Information and Software Technology,129.

[10] Rakesh Rana , Miroslaw Staron , Christian Berger , Jorgen Hansson , Martin Nilsson , Wilhelm Meding , Analyzing Defect Inflow Distribution and Applying Bayesian Inference Method for Software Defect Prediction in Large Software Projects, The Journal of Systems & Software (2016), doi: 10.1016/j.jss.2016.02.015.

[11] Guisheng Fan, Xuyang Diao, Huiqun Yu, Kang Yang and Liqiong Chen,"Software Defect Prediction via Attention-Based Recurrent Neural Network", Scientific Programming,2019.

[12] Sushant Kumar Pandey, Ravi Bhushan Mishra, Anil Kumar Tripathi,"BPDET: An effective software bug prediction model using deep representation and ensemble learning techniques",Expert Systems with Applications,144,2020.

[13] Rodrigo A. Coelho; Fabrício dos R.N. Guimarães; Ahmed A.A. Esmin," Applying Swarm Ensemble Clustering Technique for Fault Prediction Using Software Metrics", 2014 13th International Conference on Machine Learning and Applications, DOI: 10.1109/ICMLA.2014.63.

[14] Shamsul Huda, Kevin Liu, Mohamed Abdelrazek, Amani Ibrahim, Sultan Alyahya, Hmood Al-Dossari and Shafiq Ahmad," An ensemble Oversampling Model for Class Imbalance Problem in Software Defect Prediction", SPECIAL SECTION ON SOFTWARE STANDARDS AND THEIR IMPACT IN REDUCING SOFTWARE FAILURES,2018.

[15] Shanthini. A, R M Chandrasekaran, "Analyzing the Effect of Bagged Ensemble Approach for Software Fault Prediction in Class Level and Package Level Metrics", International Conference on Information Communication and Embedded Systems (ICICES2014), DOI: 10.1109/ICICES.2014.7033809.

[16] Nageswara Rao Moparthi, Dr. N. Geethanjali," Design and implementation of hybrid phase based ensemble technique for defect discovery using SDLC software metrics", 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), DOI: 10.1109/AEEICB.2016.7538287.

[17] Steven Young; Tamer Abdou; Ayse Bener," A Replication Study: Just-in-Time Defect Prediction with Ensemble Learning", 2018 IEEE/ACM 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE).

[18] Arvinder Kaur and Kamaldeep Kaur," Performance Analysis of Ensemble Learning for Predicting Defects in Open Source Software", 2014 International Conference on Advances in Computing,Communications and Informatics (ICACCI).

[19] Thanh Tung Khuat, My Hanh Le, "Ensemble learning for software fault prediction problem with imbalanced data", International Journal of Electrical and Computer Engineering (IJECE),9(4),2019.

[20] Jaroslaw Hryszko, Lech Madeyski, ” Cost Effectiveness of Software Defect Prediction in an Industrial Project”, Foundations of Computing and Decision Sciences,43(1),2018.

[21] Kazuya Tanaka; Akito Monden; Zeynep Yücel,” Prediction of Software Defects Using Automated Machine Learning”, 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD),DOI: 10.1109/SNPD.2019.8935839.

[22] Pradeep Singh,” Stacking based approach for prediction of faulty modules”, 2019 IEEE Conference on Information and Communication Technology (CICT).

[23] Haitao He, Xu Zhang, Qian Wang, Jiadong Ren, Jiaxin Liu, Xiaolin Zhao, Yongqiang Cheng,” Ensemble MultiBoost Based on RIPPER Classifier for Prediction of Imbalanced Software Defect Data”, IEEE Access,7.

[24] Z. Li, X. Jing, X. Zhu and H. Zhang, "Heterogeneous Defect Prediction Through Multiple Kernel Learning and Ensemble Learning," 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), Shanghai, 2017, pp. 91-102, doi: 10.1109/ICSME.2017.19.

[25] X. Xia, D. Lo, S. J. Pan, N. Nagappan and X. Wang, "HYDRA: Massively Compositional Model for Cross-Project Defect Prediction," in IEEE Transactions on Software Engineering, vol. 42, no. 10, pp. 977-998, 1 Oct. 2016, doi: 10.1109/TSE.2016.2543218.

[26] Boehm, B. W., Software Engineering Economics, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[27] Satya Srinivas Maddipati, Dr. G Pradeepini,Dr. A Yesubabu,” Software Defect Prediction using Adaptive Neuro Fuzzy Inference System”, International Journal of Applied Engineering Research ,ISSN 0973-4562 ,Volume 13, Number 1 (2018) pp. 394-397.

[28] R Anand, Dr. K David, Dr.S. Stanley Sagayaraj,”Identifing the impact of Defects among the Defect types in Software Development Proects”, ICSTM, May 2015,pp. 146-152.

Design of Modern Distributed Systems based on Microservices Architecture

Isak Shabani[1], Endrit Mëziu[2], Blend Berisha[3], Tonit Biba[4]
Department of Computer Engineering
Faculty of Electrical and Computer Engineering
University of Prishtina
Prishtina, Kosovo

***Abstract*—Distributed systems are very commonplace nowadays. They have seen an enormous growth in use during the past few years. The idea to design systems that are robust, scalable, reliable, secure and fault tolerance are some of the many reasons of this development and growth. Distributed systems provide a shift from traditional ways of building systems where the whole system is concentrated in a single and indivisible unit. The latest architectural changes are progressing toward what is known as microservices. The monolithic systems, which can be considered as ancestors of microservices, cannot fulfill the requirements of today's big and complex applications. In this paper we decompose a monolithic application into microservices using three different architectural patterns and draw comparisons between the two architectural styles using detailed metrics that are generated from the Apache JMeter tool. The application is created via .NET framework, uses the MVC pattern and is fictive. The two comparable apps before testing with Apache JMeter, will be deployed in almost identical hosting environment in order to gain results that are valuable. Using the generated data, we deduce the advantages and disadvantages of the two architectural styles.**

Keywords—Distributed systems; microservice; monolithic; web services; JMeter

I. INTRODUCTION

Microservices are a new development, coming into light just a few years ago. They offer many advantages compared to the old monolithic architectures. That is why many of the big tech companies have successfully made the switch to microservices. Currently, the monolithic architecture is the default model for creating a software application. Its trend is decreasing as it cannot keep up with the demands and the challenges of the new applications that are now quite big and complex.

In the monolithic architecture, application is built as a single indivisible unit. This usually means that the application has three core components that interchange information with each other: a user interface, a server-side and a database [1]. This architecture is characterized by a huge code base and has almost no modularity. Because they have a single code base, they can become so large and hence difficult to maintain. The whole application will need to be redeployed from a single small change in the code. More crucial is the fact that it is not very reliable since a bug in any part of the code can bring down the whole application [1].

Monolithic architecture, however, has some subtle advantages and with some tweaks it can still be useful to many modern applications. These include: the easiness of deployment (since only one file needs to be deployed), the easiness of development (compared to the microservices) and the network latency and security which are more noticeable in the microservices architecture. Monolithic architecture is also very easy to test. We can do so by simply launching the app and testing the UI with Selenium. However, some of the drawbacks of this architecture have made the switch to microservices a necessity [2].

Because today's apps are big and complex, in order to be useful, they need to be robust and reliable. The resources must be utilized efficiently so the users can get a seamless experience while surfing the app. Many components of the app might have different resource requirements. Some might need more CPU cycles, some others more memory etc. This imposes the need to scale the different components, independently. Scaling in the monolithic architecture is done by creating copies of the app. This means that all of these copies will access all of the data which in turn makes caching less effective and increases memory consumption and I/O traffic [2].

As authors in [3] put it, one of the problems that can arise from the monolithic applications is the evolvement into a "big ball of mud" state, which is a situation in which none of the developers understand the entire application. To overcome the obstacles, microservices provide a very reasonable and effective architectural style, which as mentioned, are increasingly being used and deployed in many modern applications. In fact, microservices are considered as the future of distributed systems.

On the other hand, despite its name, microservices are by no means, small. In this architectural style, the application is made up of a suite of small devices, all of which have their own unique codebases.

Microservices use lightweight mechanisms, somewhat like an API, to communicate between different services. Contrary to monolithic architecture, these services can be deployed together or separately. These services are loosely coupled (or headless) making this architectural style mostly decentralized [3].

It must be understood that a microservice is not a layer within a monolithic application. It has its self-contained functionalities with clear interfaces, and through its own internal components, must implement a layered architecture. According to the author in [4] this architecture follows the Unix philosophy of "do one thing and do it well". In the following sections we will explore some of the main advantages of microservices and whether it is a good idea to fully deploy an application into microservices.

The research questions we will try to answer from our experiment and analysis of literature, are

- Does decomposing into microservices impact the system's average response time?
- Is it always adequate to develop an app using the microservices logic?
- To what extent should the monolithic application be decomposed into a microservice?

A. Design and Structure of Monolithic Applications

A monolithic application describes a single-tiered software application in which the user interface and data access code are combined into a single program from a single platform. Schematically, this can be seen Fig. 1.

It is self-contained, and independent from other computing applications.

The design philosophy is that the application is responsible not just for a task but can perform every step needed to complete a particular function. Layered architecture is a common pattern seen in monolithic applications. This architecture allows for the technical capability to be changed fairly easily, especially if they are isolated to a particular layer [5].

The main idea behind this architecture is the separation of concerns, the main monolithic application components which include authorization, presentation, business logic and database are organized into four main categories or layers:

- The presentation layer contains all of the classes responsible for presenting the UI to the end-user or sending the response back to the client.
- The application layer contains all the logic that is required by the application to meet its functional requirements.
- The domain layer represents the underlying domain, mostly consisting of domain entities and, in some cases, services.
- The infrastructure layer (also known as the persistence layer) contains all the classes responsible for doing the technical stuff, like persisting the data in the database including DAOs or repositories.

An example of monolithic system architecture of real-world application is shown on Fig. 2. The diagram shows main components needed to build an Ecommerce application which authorizes costumer, takes an order, checks products inventory, authorizes payment and ships ordered products [6].

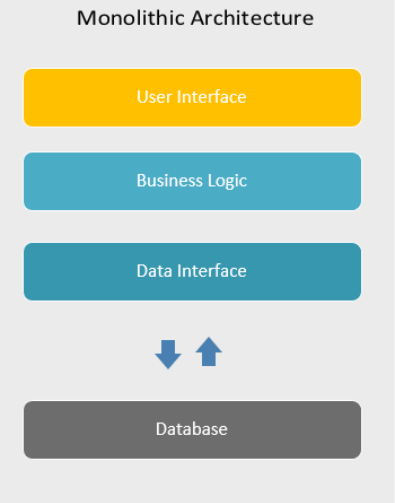


Fig. 1. Monolithic Application Architecture.

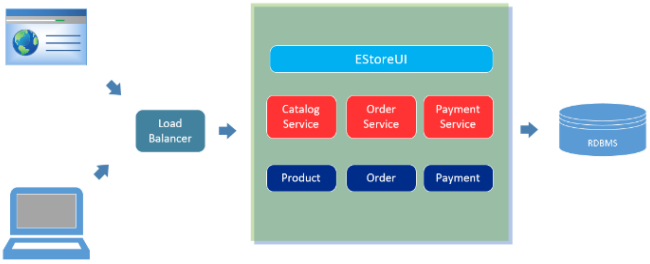


Fig. 2. Monolithic Architecture (Ecommerce Application).

Despite having many components which are independent from each other the system as shown in Fig. 2 is build and deployed as one application. With issues regarding maintenance, response time and scaling, monolithic architecture should be avoided when designing large and complex applications which may be used in different environments with different configurations or in applications which may change and need to be frequently updated.

This paper is structured as follows: Section II presents the state of the art, Section III methodology and results, Section IV case study and Section V conclusions.

II. State-of-the-Art

As mentioned previously, over the last decades, industry demands have pushed software design and architectures in various directions. The ever-growing complexity of enterprise applications, along with change and evolution management ushered in the rise of different architectures with an aim to replace or improve the traditional unified software designing model known as monolithic architecture.

Various architectures (besides the eminent ones) have been designed, researched, and used in industry, in recent years there has been a lot of hype regarding the new architectural model called microservice architecture. Considered new, microservice architecture has found itself being researched and compared a lot with existing architectures including SAO, serverless and monolithic architectures. Most of research

studies were oriented on performance analysis, cost, and resource usage.

In a research that was done by Singh and Peddoju, the performance of a monolithic application is compared to a microservices application, the applications that were built were tested for their response time and throughput. Obtained results made it clear that microservices architecture has a better performance especially when it is used for a large number of requests [7].

Similar approach was used by the IBM research team in Tokyo, they compared the performance of the monolithic and microservices applications in different environments and configurations. The results were compared for throughput, scalability, number of CPU instructions for request and number of clock cycles to complete one instruction. The results showed a significant performance boost in monolithic architecture applications in many configurations and environments, which in a way contradicts the results shown by Singh and Peddoju [8].

Microservices are often compared to Service Oriented Architecture. The research paper done by Cerny, Donahoo and Pechanec compares and analyses microservices, service-oriented architecture and self-contained systems in terms of service and architecture, characteristics, integrations, capabilities, and flexibility. The drawn conclusion presented at the end of the paper favorizes SOA for large systems with many shared components and suggests using microservices for medium distributed systems which may need to scale in the future [9].

A different approach was used on research paper done by Chen, Li and Zheng from Nanjing University. This paper discusses ways to decompose a monolithic application to microservice architecture. Throughout the paper the researchers used a top-down analysis approach and developed a dataflow-driven decomposition algorithm. They defined a three-step procedure for process decomposition involving business requirement analysis, usage of dataflow-driven algorithm and individual modules extraction [10].

According to the fourth annual Developer Ecosystem Survey conducted by JetBrains, about 85% of 19,696 developers who were surveyed in the beginning of 2020, use the microservices-based system design [11]. The programming languages of choice for building microservices are JavaScript and TypeScript; REST APIs are used for communication between microservices the most, whereas the favorite cloud provider for microservices is Amazon Web Services, as shown in [12].

Improving scalability and improving performance are two of the most important topics when it comes to microservices. In the State of Microservices 2020 research project [12], over 650 developers (CTOs, Lead Developers, and Senior Developers) were asked to rate in scale 1-5 how they enjoy working with microservices when it comes to different aspects.

As shown in [2] Table I, most experts are happy with microservices for solving scalability issues, whereas maintenance and debugging seem to be a challenge for them.

TABLE I. WORKING WITH MICROSERVICES

Category	Average rating (1-5)
Setting up a new project	3.8
Maintenance and debugging	3.4
Efficiency of work	3.9
Solving scalability issues	4.3
Solving performance issues	3.9
Teamwork	3.9

Regarding security, there are still many challenges due to the complexity of the developments, the hardness of monitoring, and debugging and auditing of the full application in foreign environments [13].

Before moving to microservices, we should be aware of the architectural challenges. Some of the main architectural challenges, as presented in [14], are:

1) Dispersed business logic – microservices approach distributes the operating logic and execution flow of complex features among many applications.

2) Lack of distributed transactions – attempting to maintain consistency among many microservices involved in business transaction is extremely complicated.

3) Inconsistent dynamic overall state – it is related to lack of distributed transactions. Overall consistency gets more complicated with data that is geographically distributed data within the same domain because of sharding and data replication.

4) Difficulty in gathering composite data – joining data for analytics of the overall system in a microservices architecture is not straightforward.

5) Difficulty in debugging failures and faults – attempting to pinpoint the source of an error might require debugging multiple applications. Identification of the root cause of the problem is difficult primarily because of deep hierarchies of microservices (AC1) and the inability to determine the exact state of the system (AC3).

6) Difficulty in evolving – software evolution is a hard concept in an environment different where parts of the system evolve continuously, in parallel.

III. METHODOLOGY AND RESULTS

This section gives an overview on which methods and tools were used.

The goal of this section is to offer a way of passing between monolithic architecture to microservices approach and comparing them. So, we are going to demonstrate how to identify key design issues of monolithic applications and how they should be reflected in microservices approach. For that purpose, we will use a monolithic application that is developed in Model-View-Controller approach, which is based on monolithic architecture, and we will try to offer a way of decomposing it in microservices approach.

We are aware that there are a lot of design patterns that exists for developing web applications. But based on usage we

have decided to use MVC as one of most used architectural patterns for developing web applications that are based on monolithic architecture and not only.

A. E-Shop Monolithic Application

As we said earlier, we will use an application that uses MVC approach, which is developed in Asp.NET Core with MVC approach. Before analyzing this application, we want to make purely understood that the term "monolithic", in this context refers to the fact that these applications are deployed as a single unit, not as a collection of interacting services and applications [15].

Application that we have developed for this paper is based on application of Microsoft [16], for e-shop. The main reason why we have chosen to develop an e-shop application is to demonstrate how to pass between monolithic to microservices is because there is an almost perfect example that microservices should be used there.

In Fig. 3, we have presented schematically controllers of the application that are developed.

As is can be seen there are four controllers that monolithic application currently has. First controller, Order, is for handling requests that are for ordering items on application. Second controller, Product, it is used for managing products. The third controller, Home, is for main and privacy terms. The last controller which is default controller for authentication and authorization is Identity, it used to manage accounts and roles. In Fig. 3 we have presented Identity as a Controller, but in latest version of Identity Microsoft uses Razor pages for this module, but we will abstract this, and we will consider as a controller.

In Fig. 4 we have presented schematically structure of application that is developed as a monolithic application in Visual studio.

As it can be seen from Fig. 4, all application logic, including presentation, business and data access logic is in one place.

In the next section we will offer a way of passing microservices approach and how we should identify parts of application that should be microservice itself.

B. Decomposing to Microservices

In this section we will try to offer a way of how to decompose E-shop application to microservices approach.

Before starting to identify microservices we want to make purely understood that there is no general method that can be applied to every monolithic application. This means that we need to study very deeply application before architecting to microservices.

For E-Shop application, the first thing that must be transformed to microservice is Identity service, which is used for authentication and authorization purpose. One the most important services in E-Shop application, and in most applications, is security. Identity service is an IdentityServer4 [17], which is a typically used for managing authentication and authorization in microservices environment. Typically, IdentityServer4 acts as a middleware [18] that adds the spec compliant OpenID Connect and OAuth 2.0. With IdentityServer4 all access to microservices can be managed and this service is responsible for generating access token for clients.

Other important microservice for E-Shop application is product microservice, which is responsible for managing products for this application. So, this microservice can register new products, edit them, or see details about products. So, this microservice does only one thing but it does in a perfect way.

Last microservice is responsible for handling orders of customers. So, this microservice is focused only on processing orders, and offers a payment for orders.

For testing purpose is developed a client which will use microservices over RESTful API [19]. A schematic presentation of E-Shop application decomposed to microservices is displayed in Fig. 5.

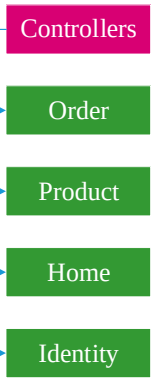


Fig. 3. Controllers for E-Shop.

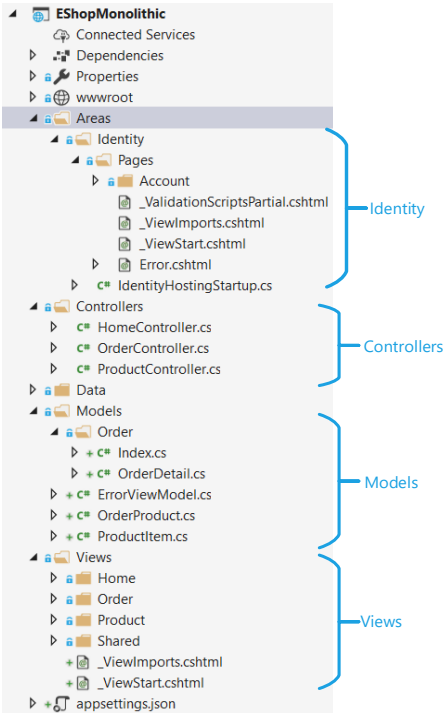


Fig. 4. E-Shop Application with Monolithic Architecture.

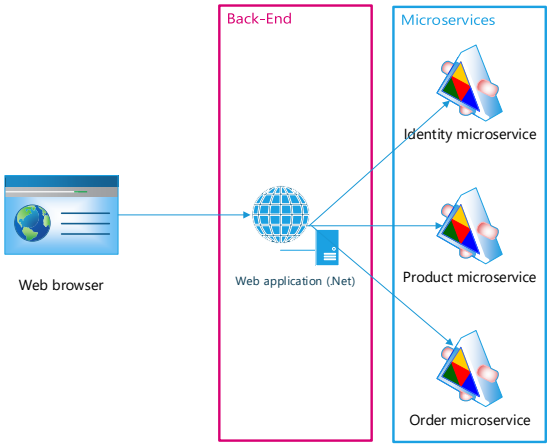


Fig. 5. E-Shop Application Decomposed to Microservices.

As it can be seen from Fig. 5, in this case we have 3 microservices, which we have described before. This decomposition offers a very good way to deal with scenarios where ordering a product is not possible, still application can offer service by listing all product that are there. So, with this decomposition we have archived a good way to handle problems with no function of order product, but order product currently contains functionality for checkout and payment. As part of comparison is this model of decomposition with E-Shop monolithic system, and other types of microservices architecture that will be presented.

As it can be seen from Fig. 5, the main problem with decomposition of E-Shop application in microservices architecture that is offered, is Order microservice, which needs to be decomposed to three microservices. These 3 microservices that will be derived from Order microservice are:

- Order microservice.
- Checkout microservice and.
- Payment microservice.

Schematically this decomposition is presented in Fig. 6.

With decomposition of Order microservice, are archived many things.

The last feature that will be applied when decomposing to microservices, in Fig. 6, is adding an API Gateway. Schematically this is presented in Fig. 7.

Decomposition that has been displayed in Fig. 7, contains an API Gateway, which acts as reverse proxy, hides functionality of microservices that are currently implemented in E-Shop application. This is a very good place to implement security for microservices.

C. Load test Comparison

In this section we will compare monolithic application with microservices for our fictive application. Comparison is made by using Apache JMeter [20] with different parameters. To have results that are comparable with each other we have hosted to Docker, with Linux container, all microservices, monolithic application and Client which consumes microservices is hosted in Internet Information services for Windows. For this purpose, we have deployed to test environment which is identic for microservices and monolithic application. Database is in Microsoft SQL server and contains same tables for both applications. Architecture of infrastructure for monolithic and microservices is presented in Fig. 8.

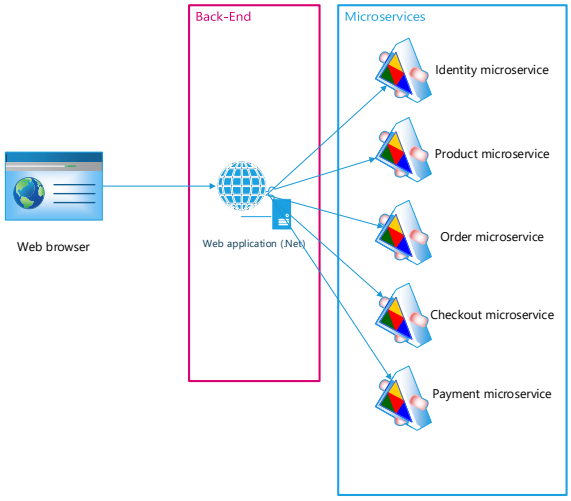


Fig. 6. E-Shop Application Decomposition Second Version.

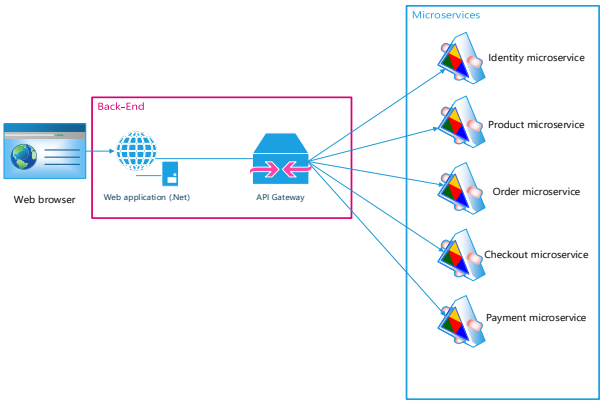


Fig. 7. Decomposition that has API Gateway.

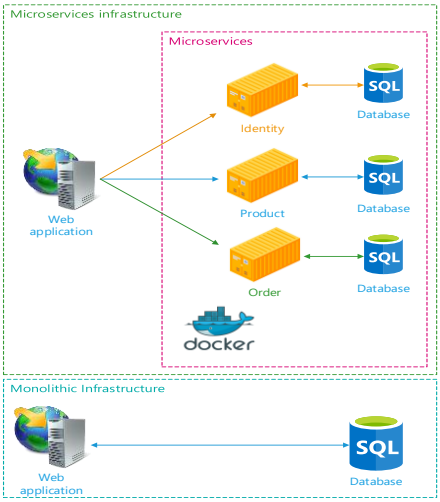


Fig. 8. Infrastructure of Microservices and Monolithic.

The first scenario will perform Get request to home page, then to list of products and finally to edit product page. All three requests are Get requests. Parameters of testing are set same for all applications. Parameters in Apache JMeter are:

- Number of Threads (users) = 100.
- Ramp-up period(seconds) = 50.
- Loop count = 5.

After creating test plan in Apache JMeter, we have gained results as can be seen in Table II.

In second comparison, as additional will be added post request which is responsible for adding new products to database. Parameters for Apache JMeter are same as above. After creating test plan in Apache JMeter, we have gained result as can be seen in Table III.

The final comparison will be made to order part. There will be added get request for checkout, order detail for specific product, update to database number of orders and finish payment.

After creating test plan in Apache JMeter, we have gained result as displayed in Table IV.

TABLE II. RESULTS FOR FIRST TEST

Parameter\Application	Monolithic	Microservices First	Microservices Second
Request	Get	Get	Get
Samples	1500	1500	1500
Average	6	10	9
Min	2	6	6
Max	41	159	98
Std. Dev.	5.14	8.83	5.33
Error %	0.00 %	0.00	0.00
Throughput	30.2/sec	10.1/sec	10.1/sec
Received KB/sec	247.14	81.38	81.53
Sent KB/sec	3.78	1.17	1.17
Avg. Bytes	8368.6	8268.6	8273.7

TABLE III. RESULTS FOR SECOND TEST

Parameter\Application	Monolithic	Microservices First	Microservices Second
Request	Get, Post	Get, Post	Get, Post
Samples	2000	2000	2000
Average	872	1851	1219
Min	2	7	8
Max	5024	7931	6361
Std. Dev.	1161.98	1858.53	1428.10
Error %	0.00 %	0.00 %	0.00 %
Throughput	22.5/sec	18.0/sec	20.7/sec
Received KB/sec	2197.87	1855.62	2154.23
Sent KB/sec	4.50	3.54	4.08
Avg. Bytes	100061.3	105852.1	106604.9

TABLE IV. RESULTS FOR THIRD TEST

Parameter\Application	Monolithic	Microservices First	Microservices Second
Request	Get, Post	Get, Post	Get, Post
Samples	2000	2000	2000
Average	7	22	21
Min	3	6	6
Max	113	127	319
Std. Dev.	5.38	15.52	16.53
Error %	0.00 %	0.05 %	0.15 %
Throughput	40.4/sec	39.8/sec	40.1/sec
Received KB/sec	220.37	3576.13	2270.08
Sent KB/sec	7.61	8.90	8.95
Avg. Bytes	5591.9	91920.7	57980.4

Very important statistic that can be derived from Table IV, is average response time that is from First and Second microservice. Decomposing to Microservices of course that has many benefits, but sometimes benefits that can be archived from decomposing might hurt performance of the system. This is proved by results displayed in Table IV.

IV. CASE STUDY

In case study will be discussed for complex system, which is implemented in Kosovo, which is Health Insurance Fund Information System of Kosovo. Because of data sensitivity we have decided to not use this system to decompose to microservices approach, so we have used a fictive application.

Results that are archived by using fictive application are very important and there can be draw parallel with Health Insurance Fund Information System and other systems.

Based on results that are archived there should be made a tradeoff between current architecture that has this system, which is monolithic application and is developed in Asp.Net, to decompose to Microservices approach. Again, based on results from Results for First Test Table II, Table III and Table IV, is evident that decomposing to microservices would decrease average response time, but benefits that could be archived from microservices, especially for this system, are bigger than the average response time. Benefits that will be archived are same as mentioned in Section C of III.

V. CONCLUSIONS

It is obvious that microservices offer a lot of advantages compared to the traditional monolithic architecture. Many of the core functionalities of microservices were described throughout the paper. Our approach in this paper, was to analyze and then compare the same application but developed with the two architectural styles. From the results obtained we saw that microservices can increase the system's average response time since there are different services that need to communicate and exchange information with one another. Testing for different parameters with Apache JMeter we saw the differences in response times between them. Results from Apache JMeter, for three cases, also told that not only response time, but also error rate is better than architecture based on microservices. On the other hand, architecture based

on microservices performs better in number of Kilo Bytes send and received per second, in case when test scenario contains post method as can be seen from Table IV.

One big advantage of microservices, is that they are not tied to a programming language. They also overcome the cumbersomeness of dealing with databases as we saw while developing our fictive application. To conclude, choosing whether to use the monolithic or the microservices architecture is not always clear cut. It all boils down to the type of application and what the developer wants to achieve. Big applications will benefit from the robustness, efficiency, and the well-organized code that the microservices make possible.

REFERENCES

[1] R. Amen, "Monolithic vs Microservices architecture," 2017.

[2] A. Kharenko, "Microservices Practioner Analysis," January 2019. [Online]. Available: https://articles.microservices.com/. [Accessed 03 June 2020].

[3] T. Jack, C. Bredley and L. Casey, "Content Stack," 03 02 2018. [Online]. Available: https://www.contentstack.com/cms-guides/decoupled-cms/monolithic-vs-microservices-cms-architectures. [Accessed 27 05 2020].

[4] K. Telai, "Medium," 09 04 2019. [Online]. Available: https://medium.com/@kenlynterai/microservices-and-distributed-systems-36a90d5d8ce. [Accessed 26 05 2020].

[5] [Online]. Available: https://medium.com/@shivendraodean/software-architecture-the-monolithic-approach-b948ded8c333. [Accessed 30 05 2020].

[6] [Online]. Available: https://medium.com/koderlabs/introduction-to-monolithic-architecture-and-microservices-architecture-b211a5955c63. [Accessed 31 05 2020].

[7] V. Singh and S. K. Peddoju. [Online]. Available: https://www.researchgate.net/publication/322001375_Container-based microservice architecture_for_cloud_applications. [Accessed 02 06 2020].

[8] T. Ueda, T. Nakaike and M. Ohara. [Online]. Available: https://dominoweb.draco.res.ibm.com/reports/RT0973.pdf. [Accessed 03 06 2020].

[9] T. Černý and M. J. Donahoo. [Online]. Available: https://www.researchgate.net/publication/320765439_Disambiguation_and_Comparison_of_SOA_Microservices_and_Self-Contained_Systems. [Accessed 05 06 2020].

[10] L. Chen, S. Li and Z. E. Li. [Online]. Available: https://www.researchgate.net/publication/323562483_From_Monolith_to_Microservices_A_Dataflow-Driven_Approach. [Accessed 07 06 2020].

[11] JetBrains, "Microservices," 2020. [Online]. Available: https://www.jetbrains.com/lp/devecosystem-2020/microservices/. [Accessed 29 January 2021].

[12] P. Mamczur, T. C. M. Mol and M. Nowak, "State of Microservices," THE SOFTWARE HOUSE, 2020.

[13] N. Mateus-Coelho, M. Cruz-Cunha and L. G. Ferreira, "Security in Microservices Architectures," in CENTRIS Conference, 2020.

[14] C. Rajasekharaiah, Cloud-Based Microservices: Techniques, Challenges, and Solutions, Suwanee: Apress, 2021.

[15] Microsoft Developer Division, .NET, and Visual Studio product teams, "Architecting Modern Web Applications with ASP.NET Core and Microsoft Azure," in Architecting Modern Web Applications with ASP.NET Core and Microsoft Azure, One Microsoft Way, 2020.

[16] Microsoft, "Github," Microsoft, 11 May 2020. [Online]. Available: https://github.com/dotnet-architecture/eShopOnWeb. [Accessed 11 June 2020].

[17] B. A. &. D. B. Revision, "IdentityServer4," [Online]. Available: https://identityserver4.readthedocs.io/en/latest/. [Accessed 1 June 2020].

[18] R. A. a. S. Smith, "Microsoft," Microsoft, 5 June 2020. [Online]. Available: https://docs.microsoft.com/en-us/aspnet/core/fundamentals/middleware/?view=aspnetcore-3.1. [Accessed 5 June 2020].

[19] E. J. R. E. R. Fielding, "ietf," ietf, June 2014. [Online]. Available: https://tools.ietf.org/html/rfc7231#section-4. [Accessed 07 June 2020].

[20] Apache JMeter, "Apache," Apache, [Online]. Available: https://jmeter.apache.org/. [Accessed 11 June 2020].

Feature Engineering for Human Activity Recognition

Basma A. Atalaa[1]*, Ibrahim Ziedan[2], Ahmed Alenany[3], Ahmed Helmi[4]
Department of Computer and Systems Engineering
Faculty of Engineering
Zagazig University, Zagazig, 44519
Egypt

***Abstract*—Human activity recognition (HAR) techniques can significantly contribute to the enhancement of health and life care systems for elderly people. These techniques, which generally operate on data collected from wearable sensors or those embedded in most smart phones, have therefore attracted increasing interest recently. In this paper, a random forest-based classifier for human activity recognition is proposed. The classifier is trained using a set of time-domain features extracted from raw sensor data after being segmented into windows of 5 seconds duration. A detailed study of model parameter selection is presented using the statistical *t*-test. Several simulation experiments are conducted on the WHARF accelerometer benchmark dataset, to compare the performance of the proposed classifier to support vector machines (SVM) and Artificial Neural Network (ANN). The proposed model shows high recognition rates for different activities in the WHARF dataset compared to other classifiers using the same set of features. Furthermore, it achieves an overall average precision of 86.1% outperforming the recognition rate of 79.1% reported in the literature using Convolution Neural Networks (CNN) for the WHARF dataset. From a practical point of view, the proposed model is simple and efficient. Therefore, it is expected to be suitable for implementation in hand-held devices such as smart phones with their limited memory and computational resources.**

Keywords*—Human activity recognition; random forest; feature engineering; sensor signal processing

I. Introduction

In daily life, a person performs diverse set of activities such as standing up, sitting down, walking, climbing stairs, etc. Automatic recognition of human activities has interesting applications in healthcare [1], keeping track of elderly people [2], and home automation [3]. Also, it has many clinical applications for stroke patients [4], Parkinson's disease patients[5], heart rate estimation [6] and in a smart health care environment [7].

The last two decades witnessed increasing interest in Human Activity Recognition (HAR) techniques due to the availability of low cost sensors specially those built-in sensors available in affordable smartphones [8-10]. Commonly used sensor types in HAR applications are accelerometers [11-14], heart rate belt sensor [15], gyroscope [16, 17], magnetometer [17], or three-inertial sensor units mounted on chest, right thigh and left ankle [12]. Such inertia devices operate at low frequencies and require low sampling rates. There are several issues which make HAR task challenging such as noisy sensor data, insufficient training examples due to few participating subjects, and the need to implement HAR systems on relatively limited-resources smart devices. Therefore, numerous studies in literature have been conducted to look for suitable representative features for activities, as well as good enough recognition models [9]. Moreover, benchmark datasets available in literature are different in type of activities, number of recorded examples for each activity, experimental settings, i.e. controlled procedure [18] whether indoor or outdoor environments [19], used sensors and sensor position on subject body. According to aforementioned factors, there is a significant variance of available HAR systems accuracy in conjunction with different datasets [20].

HAR recognition techniques can be grouped into two main categories. The first is based on computer vision [21, 22] and the second is based on data collected from one or more sensors. What makes the latter approach appealing is that sensors are affordable and are usually found in reasonably priced smartphones. Another advantage is that computational and storage requirements for processing sensor data is less than those required for image processing techniques.

In this work, the relatively challenging Wearable Human Activity Recognition Folder (WHARF) dataset is extensively investigated. This dataset is collected using a tri-axial accelerometer placed on the right wrist of subjects; hence it emulates a smart watch. It is chanllenging because of its small sampling rate, 32 Hz, compared to other datasets collected using e.g. 50 Hz sampling frequency. Real-time considerations for HAR systems require dealing with segments of data points with window length between 2 seconds and 10 seconds. Therefore, sensors with small sampling rate will deliver fewer data points complicating the task of HAR system. Moreover, there are 12 different activities in WHARF with few number of examples per activity [13]. The proposed approach here applies data preporcessing in which signals are filtered using a low-pass filter and then scaled so that all features lie within the same range. In the second step, data is segmented into windows of length 5 seconds with 50% overlapping. In the third step, several effective time-domain functions or features are extracted. The proposed classifier employs the Random Forest (RF) algorithm which achieves the best precision and also the best training time compared to other classifiers such as Artificial Neural Networks (ANN) and Support Vector Machine (SVM). The proposed system is expected to be efficient and resource-friendly for smart devices. Besides, sensitivity analysis of proposed system components such as RF parameters, some important features and preprocessing scaling step is conducted. Also, feature importance is discussed using the statistical t-test.

*Corresponding Author

The contribution of this work can be highlighted as follows: (1) introducing RF-based effective and efficient HAR system with average precision of 86.1% and average accuracy of 84.8% which improves the state-of-the-art rate of 79.1% for WHARF dataset, (2) testing the proposed system on the challenging WHARF datase which is considered in only few studies in literature [23] and [24], (3) discussing the practical implementation issues of proposed system which is important in case of further system application on smart devices, and (4) conducting sensitivity analysis of important system components to determine the optimal settings for proposed system.

The rest of this paper is organized as follows. In Section II, relevant related work in the literature is reviewed. The set of features to be employed and the proposed Random Forest-based classifier are presented in Sections III and IV, respectively. In Section V, a set of experiments are conducted to evaluate the performance of the proposed model and compare it to other machine learning techniques. Sensitivity analysis is preformed to optimally select the parameters of the proposed model in Section VI. Finally, conclusions and possible future work are drawn in Section VII.

II. Related Work

The HAR procedure from preprocessed raw sensory data can be divided into two steps: (1) extracting relevant key features from collected data signals (so-called feature engineering), and (2) classifying the observed activity based on the extracted features. The reduction of data dimensionality may can also be required using e.g. principle component analysis [25]. Due to the diversity of feature types and the classifiers that can be used in these two steps, respectively, the literature of HAR problem is wide and extensive.

Sensors such as tri-axial accelerometer and gyroscope provide time domain acceleration and angular velocity readings in the *x*, *y*, and *z* axes, respectively. In the literature, the various types of features which are extracted from such raw data can be divided into two categories:

1) Time domain features: e.g. the coefficients of an autoregressive (AR) model for each of the *x*, *y*, and *z* axes [11, 18, 26-29], signal magnitude area (SMA) [11, 18, 26-28, 30], tilt angle [11, 31], Histogram [17], mean [17, 26, 31], standard deviation [25, 26], Jerk [32, 33], roll angle [11, 24] skewness, kurtosis and total integral of modulus of accelerations (IMA) [12], and.

2) Frequency domain features: e.g. power spectral density (PSD) [12, 25], signal entropy and spectral energy [12, 31], largest frequency component, average frequency signal skewness, and frequency signal kurtosis [26].

It should be noted that the use of various types of features is important to improve the classification task. Each class of activities has its own set of discriminative features which is in general different from other classes. For example, the standard deviation feature can be used to distinguish between static and dynamic activities, and the Fast Fourier transform (FFT) coefficients can be used to distinguish between walking and running [11].

On the other hand, classifiers used in HAR studies can be classified into supervised or unsupervised. Supervised classifiers [20] include multilayer neural networks [17, 18, 30, 31, 34], support vector machine (SVM) [11, 12], decision trees [30, 31], random forest [12], k-Nearest Neighbors (kNN) [12, 16] and Bayes classifier [16, 25]. Unsupervised technique, on the other hand, include Gaussian mixture model (GMM) [13], linear-discriminant analysis [27, 28], minimal learning machine (MLM) [16], *k*-means clustering, convolutional neural networks (CNN) [35-37] and hidden Markov model (HMM) [12].

III. Time-Domain and Statistical Features

In this section, the set of features extracted from pre-processed raw acceleration signals is listed. It is assumed that there is a three-dimensional dataset of size N data points collected from an accelerometer or a gyroscope, $a_x(i)$, $a_y(i)$, $a_z(i)$, $i = 1, 2, \cdots, N$, for the x, y, and z dimensions. The data is first filtered using low pass filter to reduce noise and extract the body acceleration $b_x(i)$, $b_y(i)$, $b_z(i)$ and gravity acceleration $g_x(i)$, $g_y(i)$, $g_z(i)$ components [24].

The set of features to be employed in classification are derived from both body and gravity acceleration signals as listed in Table I. The body acceleration signal features include the mean (M) and standard deviation (STD) of filtered signals, autoregressive model coefficients, signal magnitude area, tilt angle, mean, standard deviation, entropy of jerk of signals, mean, standard deviation, power and entropy of jerk of roll angle. For gravity acceleration component, the signal power along each axis and the mean of angle of *x*-axis component are used.

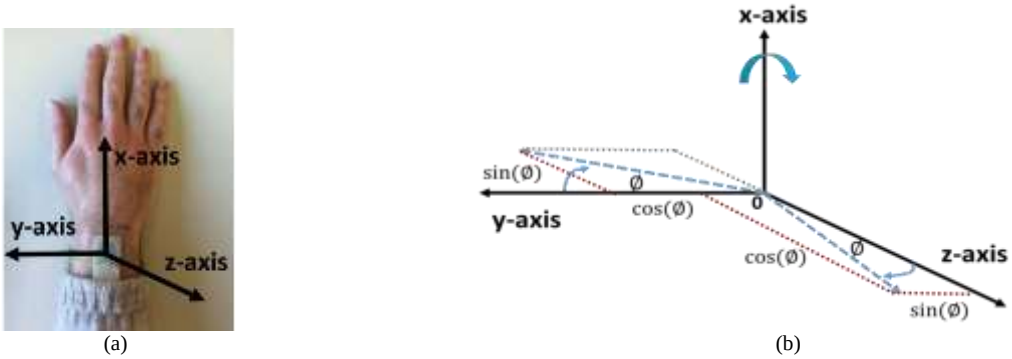
IV. The Proposed Model

The proposed classifier consists of three stages as shown in Fig. 2. In the first stage, the data is applied to a low pass filter to filter out noise and separate body acceleration from gravity acceleration. The data is then segmented into windows of 5 seconds duration consisting of 160 data points. In the second stage, the set of features listed in Table I are extracted. Finally, the classification task is performed in the third stage using random forest classifier [12].

Random Forest can be described as an ensemble or set of decision trees as shown in Fig. 2 where each tree produces a prediction of the class to which the given example belongs. The overall decision is then made using a voting process on the most predicted class among all trees in the forest. Random forest classifier has several so-called hyper-parameters which affect the classification. These include the number of trees in the forest and the maximum depth of the trees. The default value for number of trees is 100 whereas the default value for the maximum depth is 0. This means that each tree will expand until every leaf is pure, i.e. all data on the leaf comes from the same class. Random Forest classifier first selects random feature vectors from the dataset, builds a decision tree for each sample and performs a vote to determine the most voted prediction. In the current work, the basic RF classifier is employed in HAR recognition. To find the optimal RF parameters, a sensitivity analysis is conducted in Section

TABLE I. LIST OF FEATURES AND THEIR FORMULAS

Term	Meaning	Formula	Scaling factor
Autoregressive (AR) model coefficients	Autoregressive model is used to predict time series data from past data records in *x, y and z*-directions	$b_x(n) = -\sum_{k=1}^{p} \mathrm{a(k)}\, b_x(n-k) + \mathrm{e(n)}$	$1/\sqrt{\| b_x \|}$
Signal magnitude area	A scalar feature used to distinguish static from dynamic activities such as standing and walking [11]	$\mathrm{SMA} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\|b_x(\mathrm{i})\| + \|b_y(i)\| + \|b_z(i)\|\right)$	$1/\|\|(b_x)^2\|\|$
Tilt Angle	Angle between *z*-axis and gravitational vector *g*. It is used to distinguish postures such as standing and lying [11]	$\Phi = \frac{1}{N}\sum_{i=1}^{N} \arcsin(\frac{b_z(i)}{\|\|b_z\|\|}$	$1 \times \sqrt{\| b_x \|}$
Jerk	The rate of change of body acceleration.	$J = \frac{\partial b_x(i)}{\partial t}$	-
Roll angle	Describes the rotation of accelerometer attached to the participant's hand about x-axis as shown in Fig. 1 [24]	$\emptyset(\boldsymbol{i}) = \boldsymbol{tan^{-1}}(-b_{\boldsymbol{z}}(\boldsymbol{i}), -b_{\boldsymbol{y}}(\boldsymbol{i}))$	$1/\sqrt{\| b_x \|}$
Angle of x-axis gravity signal	This angle is used to estimate sensor attitude	$\theta = real\left(\cos^{-1}\left(\max(\min(\frac{u^T v}{\|\|u\|\|\|\|v\|\|}, 1), -1)\right)\right)$	$1/\| b_x \|$
Power	Signal power	$P_x = \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(g_x(i)^2)}\right)^2$	-
Entropy of signal (S)	Statistical measure of signal randomness	$E(S) = -\sum_{i=1}^{n}(P(s_i)\, log_2\, P(s_i))$	-
Mean	Describes the central tendency or the dc level of the signal	$\mu_x = \frac{1}{N}\sum_{i=1}^{N} b_x(i)$	$1/\sqrt{\| b_x \|}$ for b_x, b_y, b_z and jerk
Standard deviation	Describes the amount of variation around the mean	$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(b_x(i) - \mu_x)^2}$	$1/\sqrt{\| b_x \|}$ for b_x, b_y, b_z



(a) (b)

Fig. 1. Accelerometer Orientation during WHARF Dataset Collection [23] and (b) Roll Angle (∅) after Rotation Around x-axis.

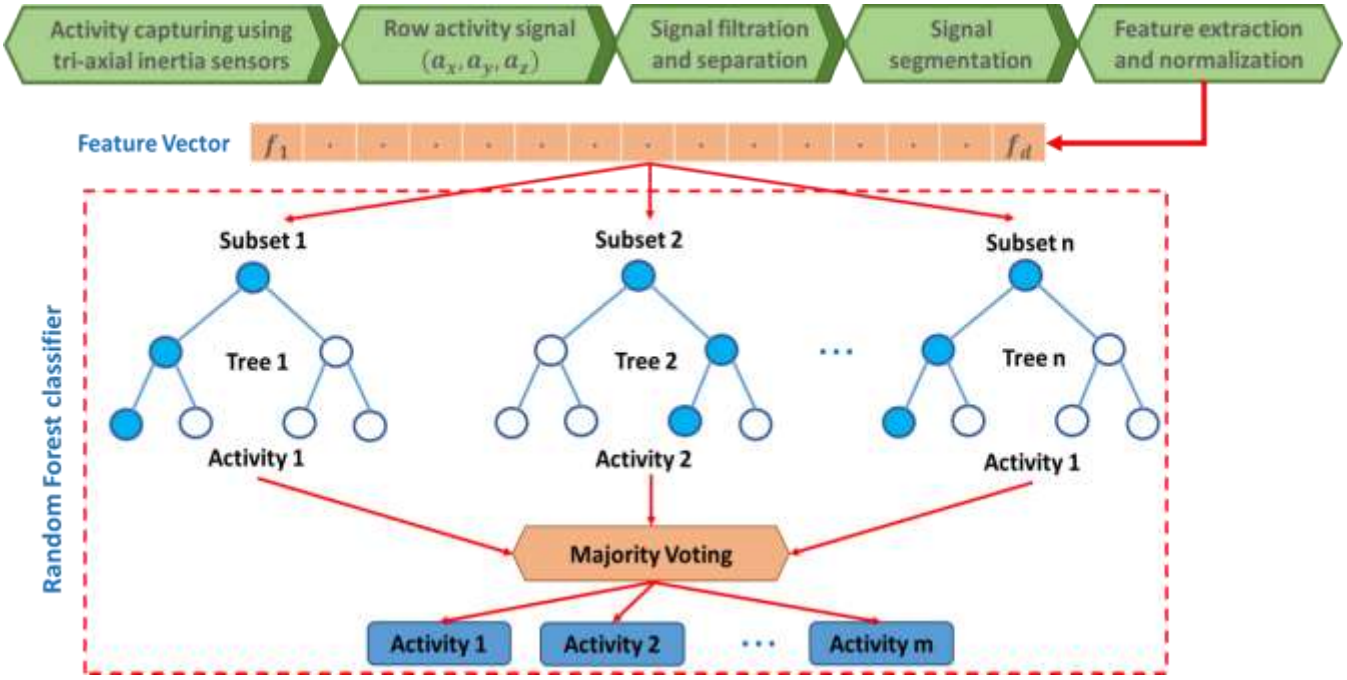


Fig. 2. Block Diagram of the Proposed Human Activity Recognition System.

V. Experimental Results

A. Dataset

In this section, the benchmark Wearable Human Activity Recognition Folder (WHARF) dataset by Bruno et al. [13], is used to examine the performance of the proposed HAR technique. The dataset was collected by an ad-hoc tri-axial accelerometer sensor attached to the right wrist of the participant. The participants are 17 volunteers; 11 males, with age ranging from 19 to 81 years; and 6 females, with ages between 56 and 85 years [11]. The digital resolution of the sensor is 6 bits and the sampling rate is 32 Hz. The dataset contains the following 12 activities: Brush_teeth (BT), Climb_stairs (CS), Comb_hair (CH), Descend_stairs (DS), Drink_glass (DG), Getup_bed (GB), Liedown_bed (LB), Pour_water (PW), Sitdown_chair (SD), Standup_chair (SU), Use_telephone (UT) and Walk (WK). The examples of each activity class are contained in a separate folder and raw signals for each single activity are saved in one text file.

B. Classification Rates

According to recent studies in the literature [23, 26, 35], classification results of different classifiers and settings have been reported in terms of the Precision (or positive predictive rate) and the Recall (or sensitivity) as the most crucial metrics in HAR applications. Let TP, FP and FN denote true positive, false positive and false negative, respectively, then the precision (P) can be calculated as $P = \frac{TP}{TP+FP}$, whereas the recall (R) is expressed as $R = \frac{TP}{TP+FN}$.

All experiments were conducted using machine learning package Sklearn in Python. Each activity signal is segmented into windows of 5 seconds duration [24] in order to fulfil real-world demands of HAR systems [26]. In Table II, a comparison is made between the proposed model using random forest against SVM and ANN. The results show that SVM and ANN have better precision than random forest in some activities. For example, SVM achieves 92.1% for Walking while ANN achieves 97% for Descend_stairs activity. However, the proposed model outperforms both SVM and ANN in terms of the average precision achieving 86.1% over all activities.

TABLE II. Comparison of Three Classifiers using the Same Feature Set in Terms of Precision Metric (%). The Activities are Brush_Teeth (BT), Climb_Stairs (CS), Comb_Hair (CH), Descend_Stairs (DS), Drink_Glass (DG), Getup_Bed (GB), Liedown_Bed (LB), Pour_Water (PW), Sitdown_Chair (SD), Standup_Chair (SU) , Use_Telephone(UT) and Walk (WK)

	BT	CS	CH	DS	DG	GB	LB	PW	SD	SU	UT	WK	Av. Pre.
SVM	83.1	73.8	86.3	87.8	85.3	66.4	46.2	83.6	75.6	65.4	97.3	92.1	78.6
ANN	92	74.3	96.9	97	88.2	63.8	68.4	79.2	79.2	64.2	82.6	82.4	80.7
RF	94.6	85	91	94.1	90.7	75.2	72.2	81.6	88.8	85.1	92.7	82.4	86.1

VI. Sensitivity Analysis and Discussion

A. RF Hyper-Parameters

The hyper-parameters of a random forest, number of trees and maximum tree depth, has a significant effect on the performance of the proposed classifier. To determine the optimal values for these parameters, the classifier is extensively tested using different sets of parameters to obtain the best possible precision. The results of this experiment is shown in Fig. 3 where it can be noticed that as the number of estimators increases, the size of the model on the disk significantly increases, however, without significant increase in the precision. Therefore, the best precision obtained is 86.1% with 100 trees and maximum depth 20. The model size on the disk, in this case, is 16 MB which is reasonable.

B. Effect of Feature Scaling and Normalization

In the proposed model, the use of normalized (scaled) features leads to an average precision of 86.1% as shown in the first row in Table III. Using un-normalized features, however, reduces the average precision to 83.5% as shown in the second row in Table III. This emphasizes the importance of feature normalization.

C. Feature Reduction based on the T-Test

It is important to check the validity and strength of features independently of the classifier to be used afterwards. This simplifies the analysis of the model and reduces the overhead of re-running the whole model several times to check the effect of every feature on the performance.

To evaluate the power of a given feature f_i in discriminating between two classes, the following *t*-test formula can be used

$$t(f_i) = \frac{|\ \mu_{i1} - \mu_{i2}\ |}{\sqrt{\frac{{\sigma_{i1}}^2}{s_1} + \frac{{\sigma_{i2}}^2}{s_2}}}$$

where f_i denotes the *i*-th feature, $\mu_{i1}$and μ_{i2} are the sample means, σ_{i1} and σ_{i2} are the sample standard deviations, and s_1 and s_2 are the size of the two classes, respectively.

Fig. 4 shows the percentage of effective features in each model. For each feature column, a two-sample *t*-test was carried out between one independent activity class and the other classes in each dataset, is a so-called *one-versus-all* binary classification. It is possible to conclude that if average *t_val* is less than or equal to a critical threshold of 10, the feature is not discriminative enough and could be safely eliminated. It is found that the average *t*-value for 13 features is less than or equal to 10 and, hence, the size of feature vector reduces to only 24 features. The previous experiments are repeated using the reduced feature vector and the results are shown in Table III. It can be seen from Table III that using the reduced set of features, the average precision decreases from 84%, using full set of features, to only 82% which may not be acceptable.

Table III also shows that the precision is high for some classes and low for others. As can be noticed in Table III, for the activities which made by hand, such as Brush_Teeth, Comb_Hair, Drink_Glass and Use_Telephone, good precision is obtained. Recalling that the sensor is attached to the wrist of the right hand, it can be realized that the position of the sensor helps in capturing these activities in a better way.

D. Size on the Disk and Training Time

In this part, we compare the proposed model using random forest to ANN and SVM in terms of model size in memory and training, and inference times. The results are shown in Table IV where it can be seen that the proposed model is superior to SVM in terms of training time. The proposed model is even better compared to ANN, although the difference is not significant if the number of iterations used in training ANN is reduced. On the other hand, the size of the proposed random forest classifier, 16 MB, is large compared to the other two classifiers and its inference is slower. The large size of the RF classifier is due to the large number of trees, 100, employed. Although the proposed model has large size and slower inference time, which is only 0.01 seconds, they are still reasonable. This combined with fast training time, makes the proposed classifier suitable for use in smartphones and hand held devices.

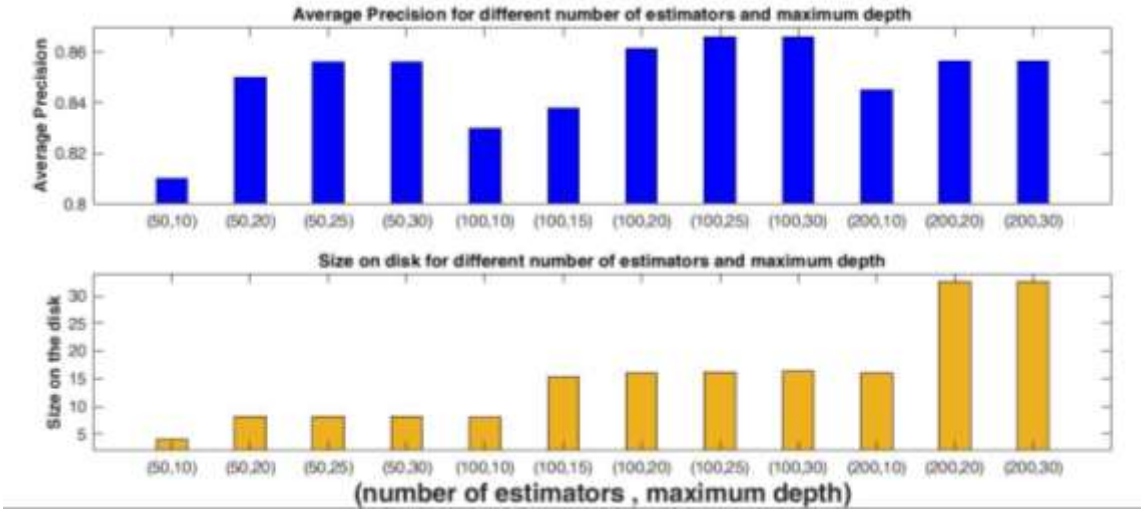


Fig. 3. Average Precision and Size on Disk for Several Combinations of the Number of Trees (Estimators) and the Maximum Depth in the RF.

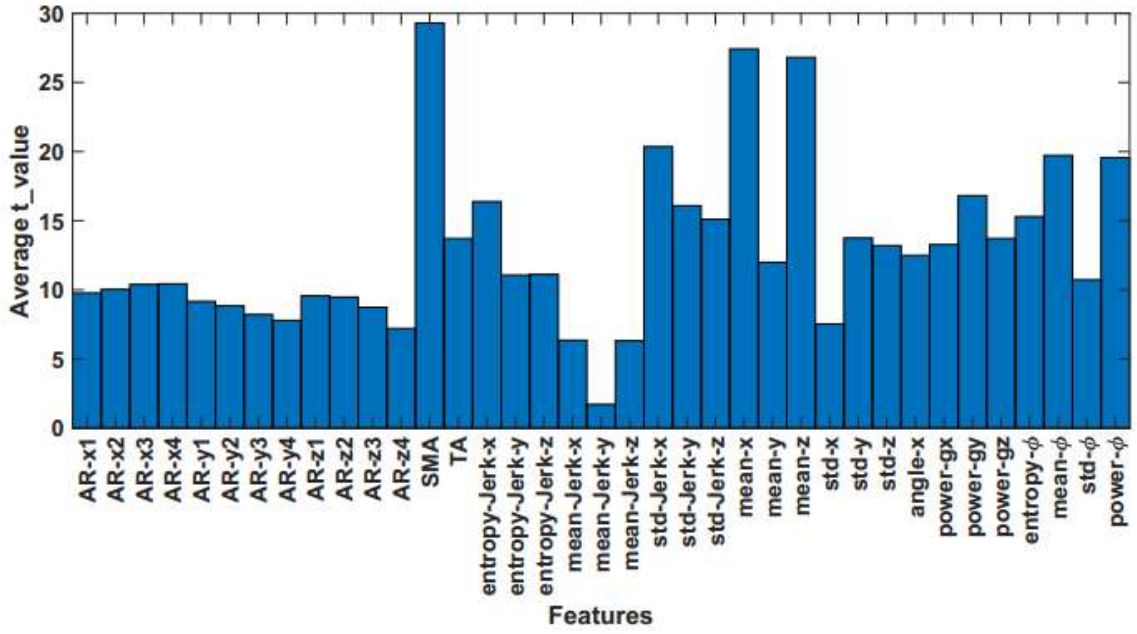


Fig. 4. Estimated Discriminating Power of Features using T-Test.

TABLE III. PRECISION (P) FOR PROPOSED MODEL USING FULL SET OF 37 FEATURES, FEATURES WITH NO SCALING AND REDUCED SET OF 24 FEATURES FOR EACH ACTIVITY CLASS IN WHARF DATASET FOR RANDOM FOREST WITH 100 ESTIMATORS AND DEPTH 20

Feature Settings	**BT**	**CS**	**CH**	**DS**	**DG**	**GB**	**LB**	**PW**	**SD**	**SU**	**UT**	**WK**	**Av. Pre. (%)**
Model feature set	94.62	85	91.04	94.12	90.76	75.17	72.22	81.62	88.89	85.11	92.68	82.39	86.1
No scaling	89.79	85.58	92.75	94.59	79.26	74.81	80	72.87	78	80	85.36	89.24	83.5
Reduced Features	94.38	86.27	87.5	89.74	83.19	62.89	100	76.69	66.67	64.06	88.09	87.03	82.2

TABLE IV. SIZE ON DISK, TRAINING AND INFERENCE TIME FOR PROPOSED RF, ANN AND SVM CLASSIFIERS USING FULL 37 SET OF NORMALIZED FEATURES

	Size on the disk	**Training time (seconds)**	**Inference time (seconds)**
RF (100 estimators, depth 20)	16 MB	1.85	.011
ANN (55 hidden neuron)	113 KB	18 (1000 iterations) 12 (500 iterations) 2.88 (100 iterations)	.001
SVM (Grid search training)	930 KB	150 (via grid search)	< .000001 sec

E. Comparison with other Studies on WHARF Dataset

In this section, we compare the proposed model to other models on the same dataset [24] and [23]. First, Jordao [24] used the same activities used in the current model. However, the author first performed data augmentation and calculated the attitude estimation as features to improve the convolution neural network performance. On the other hand, Aguirre [23] dealt with the raw data, performed feature extraction and then introduced the features to a SVM classifier. In the aforementioned studies, accuracy the model has been reported, thus for the proposed model here average accuracy is calculated as the ratio of total number of correctly predicted labels to total number of tested labels. Results of classification accuracy shown in Table V reveal the superiority of proposed model.

F. Limitations of the Current Work

The proposed model (features + classifier) is tested only on one dataset. However, in order to well investigate the generalization of such model, there is a need to test more benchmark datasets for human activity recognition. In addition, the model needs to be refined in order to achieve real time requirements such as considering smaller window size. Also, current study lacks to consider the effect of variant sampling rates of employed sensors. For WHARF data set, sensor is 32 Hz whereas sensors embedded in smartphones are usually 50 Hz. Similarly, sensor or device orientation is expected to affect such HAR models performance. Here, the effect of roll angle is considered, however other dynamic movements of human limbs (i.e. wrist, shoulder, waist or leg) are vital for determining the most suitable feature set.

TABLE V. COMPARISON OF THE PERFORMANCE OF THE PROPOSED MODEL TO PREVIOUS STUDIES [24] AND [23] ON THE WHARF DATASET

	Feature Extraction	Features domain	Classifier	Average Accuracy (%)
Aguirre [23]	Engineered	Time-domain	SVM	66.48
Jordao [24]	Raw acceleration signals	Activations of convolution layers	CNN	79.31
Proposed model	Engineered	Time-domain	RF	84.86

VII. CONCLUSION AND FUTURE WORK

In this work, a simple classification model based on random forest classifier has been proposed for human activity recognition tasks. HAR becomes a very attractive field not only due to the wide range of applicability of machine learning tools, but also for important applications like rehabilitation, health monitoring and clinical applications. The proposed technique employs a feature vector consists of several time-domain features extracted from accelerometer sensor data such as AR model coefficients, mean, and standard deviation. The proposed model is shown to achieve better average accuracy compared to other methods proposed in the literature such as SVM and ANN. RF also has a better classification rate compared to CNN on the same WHARF dataset. The proposed system was trained for segmented data as done in some previous studies.

Examining the implementation of the proposed model on smart devices can be examined in future work. It also possible to use more than one sensor embedded in smartphones instead of using one wearable sensor as in WHARF. This opens a window for an interesting extension in HAR field concerning implementation of efficient and accurate models on personal devices and examining them in practical environments. Another reasonable extension for this work is to deal with signals that may contain readings of more than one activity. For example, the user may be in a continuous movement where he or she switches between some activities like walking, climbing stairs, sitting and others. It is therefore interesting to examine the performance of the proposed models in literature in real-time situations and ensure that they achieve results similar to those obtained off line.

REFERENCES

[1] X. Liu, L. Liu, S. J. Simske, and J. Liu, "Human daily activity recognition for healthcare using wearable and visual sensing data," in 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016, pp. 24-31.

[2] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," International Journal of Distributed Sensor Networks, vol. 12, p. 1550147716665520, 2016.

[3] A. Wang, G. Chen, C. Shang, M. Zhang, and L. Liu, "Human activity recognition in a smart home environment with stacked denoising autoencoders," in International conference on web-age information management, 2016, pp. 29-40.

[4] C. Jobanputra, J. Bavishi, and N. Doshi, "Human Activity Recognition: A Survey," Procedia Computer Science, vol. 155, pp. 698-703, 2019.

[5] W. Sousa Lima, E. Souto, K. El-Khatib, R. Jalali, and J. Gama, "Human Activity Recognition Using Inertial Sensors in a Smartphone: An Overview," Sensors, vol. 19, p. 3213, 2019.

[6] W. Qi, H. Su, and A. Aliverti, "A Smartphone-Based Adaptive Recognition and Real-Time Monitoring System for Human Activities," IEEE Transactions on Human-Machine Systems, 2020.

[7] P. L. Aguirre, L. A. Torres, and A. P. Lemos, "Autoregressive modeling of wrist attitude for feature enrichment in human activity recognition," in Congresso Brasileiro de Inteligência Computacional, 2017.

[8] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," Sensors, vol. 15, pp. 31314-31338, 2015.

[9] B. Bruno, F. Mastrogiovanni, and A. Sgorbissa, "A public domain dataset for ADL recognition using wrist-placed accelerometers," in the 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014, pp. 738-743.

[10] H. Zhang, Z. Xiao, J. Wang, F. Li, and E. Szczerbicki, "A Novel IoT-Perceptive Human Activity Recognition (HAR) Approach Using Multihead Convolutional Attention," IEEE Internet of Things Journal, vol. 7, pp. 1072-1080, 2019.

[11] S. Oniga and J. Sütő, "Human activity recognition using neural networks," in Proceedings of the 2014 15th International Carpathian Control Conference (ICCC), 2014, pp. 403-406.

[12] L. B. Marinho, A. H. de Souza Júnior, and P. P. Rebouças Filho, "A new approach to human activity recognition using machine learning techniques," in International Conference on Intelligent Systems Design and Applications, 2016, pp. 529-538.

[13] R.-A. Voicu, C. Dobre, L. Bajenaru, and R.-I. Ciobanu, "Human physical activity recognition using smartphone sensors," Sensors, vol. 19, p. 458, 2019.

[14] A. M. Khan, Y.-K. Lee, and T.-S. Kim, "Accelerometer signal-based human activity recognition using augmented autoregressive model coefficients and artificial neural nets," in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, pp. 5172-5175.

[15] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla, "Sensor-based datasets for human activity recognition–a systematic review of literature," IEEE Access, vol. 6, pp. 59192-59210, 2018.

[16] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human Activity Recognition using Inertial, Physiological and Environmental Sensors: a Comprehensive Survey," arXiv preprint arXiv:2004.08821, 2020.

[17] N. Capela, E. Lemaire, N. Baddour, M. Rudolf, N. Goljar, and H. Burger, "Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants," Journal of neuroengineering and rehabilitation, vol. 13, pp. 1-10, 2016.

[18] W.-Y. Cheng, A. Scotland, F. Lipsmeier, T. Kilchenmann, L. Jin, J. Schjodt-Eriksen, et al., "Human activity recognition from sensor-based large-scale continuous monitoring of Parkinson's disease patients," in 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017, pp. 249-250.

[19] E. Brophy, W. Muehlhausen, A. F. Smeaton, and T. E. Ward, "Optimised Convolutional Neural Networks for Heart Rate Estimation and Human Activity Recognition in Wrist Worn Sensing Applications," arXiv preprint arXiv:2004.00505, 2020.

[20] A. Subasi, K. Khateeb, T. Brahimi, and A. Sarirete, "Human activity recognition using machine learning methods in a smart healthcare environment," in Innovation in Health Informatics, ed: Elsevier, 2020, pp. 123-144.

[21] M. M. Moussa, E. Hamayed, M. B. Fayek, and H. A. El Nemr, "An enhanced method for human action recognition," Journal of advanced research, vol. 6, pp. 163-169, 2015.

[22] R. Poppe, "A survey on vision-based human action recognition," Image and vision computing, vol. 28, pp. 976-990, 2010.

[23] X. Long, B. Yin, and R. M. Aarts, "Single-accelerometer-based daily physical activity classification," in 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009, pp. 6107-6110.

[24] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in Esann, 2013.

[25] M.-W. Lee, A. M. Khan, and T.-S. Kim, "A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation," Personal and Ubiquitous Computing, vol. 15, pp. 887-898, 2011.

[26] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim, "A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer," IEEE transactions on information technology in biomedicine, vol. 14, pp. 1166-1172, 2010.

[27] A. M. Khan, Y.-K. Lee, S.-Y. Lee, and T.-S. Kim, "Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis," in 2010 5th international conference on future information technology, 2010, pp. 1-6.

[28] G. Krassnig, D. Tantinger, C. Hofmann, T. Wittenberg, and M. Struck, "User-friendly system for recognition of activities with an accelerometer," in 2010 4th International Conference on Pervasive Computing Technologies for Healthcare, 2010, pp. 1-8.

[29] A. A. Sukor, A. Zakaria, and N. A. Rahim, "Activity recognition using accelerometer sensor and machine learning classifiers," in 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), 2018, pp. 233-238.

[30] W. Hamäläinen, M. Järvinen, P. Martiskainen, and J. Mononen, "Jerk-based feature extraction for robust activity recognition from acceleration data," in 2011 11th International Conference on Intelligent Systems Design and Applications, 2011, pp. 831-836.

[31] M. Ghobadi, J. Sosnoff, T. Kesavadas, and E. T. Esfahani, "Using mini minimum jerk model for human activity classification in home-based monitoring," in 2015 IEEE International Conference on Rehabilitation Robotics (ICORR), 2015, pp. 909-912.

[32] A. Jordao, L. A. B. Torres, and W. R. Schwartz, "Novel approaches to human activity recognition based on accelerometer data," Signal, Image and Video Processing, vol. 12, pp. 1387-1394, 2018.

[33] M. Altuve, P. Lizarazo, and J. Villamizar, "Human activity recognition using improved complete ensemble EMD with adaptive noise and long short-term memory neural networks," Biocybernetics and Biomedical Engineering, 2020.

[34] K. Xia, J. Huang, and H. Wang, "LSTM-CNN Architecture for Human Activity Recognition," IEEE Access, vol. 8, pp. 56855-56866, 2020.

[35] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," Expert systems with applications, vol. 59, pp. 235-244, 2016.

[36] E. Kim, "Interpretable and Accurate Convolutional Neural Networks for Human Activity Recognition," IEEE Transactions on Industrial Informatics, 2020.

[37] P. L. R. Aguirre, "Feature Enrichment in Human Activity Recognition."

Digitization of Supply Chains as a Lever for Controlling Cash Flow Bullwhip: A Systematic Literature Review

Hicham Lamzaouek[1], Hicham Drissi[2], Naima El Haoud[3]
ENCG Casablanca/ Hassan 2 University, Casablanca, Morocco

***Abstract*—Due to the new possibilities offered by digital technologies, more and more companies are embarking on a process of digitizing their supply chains. This dynamic seems to be the opportunity to analyse the impact that digital technologies may have on one of the phenomena that disrupt financial flows within supply chains, and that can alter the companies' treasury, namely that of cash flow bullwhip (CFB). The results of the systematic literature review that was carried out allow to affirm that several technologies can contribute positively to limiting this phenomenon and this by acting on these operational causes, which are the reliability of forecasts, batch orders, the fluctuation in sales prices, rationing games, and lead times.**

***Keywords*—Cash flow bullwhip; digital technologies; digitization; supply chain; cash flow; bullwhip effect**

I. Introduction

Without controlling its financial flows, the company cannot continue to operate even if it proposes an attractive offer to the market [1]. In this regard, it is obvious that any company must pay particular attention to this part of the flows which are in reality only the consequences of the physical and information flows of a supply chain. In fact the overall performance of companies is linked with their financial performance [2], [3], which make the control of financial flows very important. That said, research has shown that these financial flows are usually undergoing a disruption, referred to as cash flow bullwhip (CFB), and which is observed in almost all supply chains [4]. In this regard, it has been shown that CFB originates from the phenomenon of the bullwhip effect [4]. Indeed, the amplification of stocks caused by the BWE, slows down the time necessary to convert them in sales, and even makes this time instable. This leads to a delay in the cash flow generation [4].

The diagram below illustrates the dynamics of the CFB at the level of the supply chain (Fig. 1).

Mathematical modeling of the CFB shows that this disturbance is caused by the bullwhip effect and supply chains' lead times [4]. This is to say that the causes of CFB are lead times, reliability of sales forecasts (quality of demand signals processing), batch orders, sales price fluctuations, and rationing games between supply chain members [5].

While several studies have tried to propose solutions for the causes listed above [6], the rise of digital technologies prompts to be interested in the possible contributions of these technologies in reducing the causes of CFB. In what follows, we propose a systematic review of the literature concerning the contribution of digital technologies to the control of CFB. We propose to structure the rest of the document in three main parts. The first part will be devoted to the research methodology. The second will present the descriptive analysis of the results. The third part will synthesize the results of the content analysis by identifying the concept of digitization and the main digital technologies, and by illustrating the impacts of these technologies on each of the operational causes of CFB. We will conclude with a summary and avenues for further research.

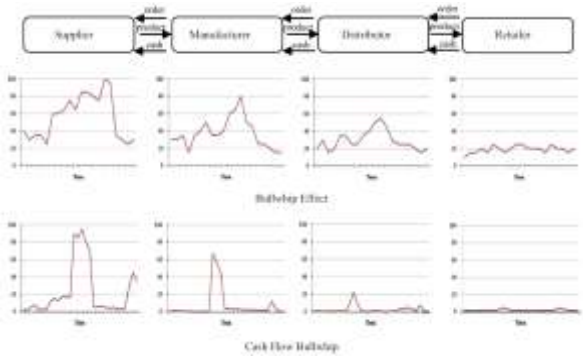


Fig. 1. BWE and CFB in Supply Chains (Tangsucheeva & Prabhu, 2014).

II. Methodology

In order to fully understand the potential impacts of digitalization on the CFB, we opt for a systematic literature review. This type of literature review is more methodical compared to narrative reviews, and establishes an in-depth description of the steps taken to select, examine and analyze relevant sources with the aim of minimizing bias and increasing transparency.

We choose the approach of Denyer and Tranfield which distinguishes the following four steps (Fig. 2):

Fig. 2. Research Methodology.

The research question can be formulated as following:

- How the use of digital technologies by supply chain members, can lead to a control of the cash flow bullwhip?

To answer this question, we need to first address the following questions:

- What is the meaning of digitization?
- What are the main digital technologies related to supply chain management?
- In which manner can the digital technologies act on the operational causes of cash flow bullwhip?

A. Definition of Keywords and Databases

Articles potentially related to our topic were identified in the "Scopus" and "Taylor & Francis" databases by using the combination of the following terms: "Digital technologies and cash flow bullwhip", "Digital technologies and bullwhip effect", "Digitization and cash flow bullwhip", "Digitization and bullwhip effect".

B. Definition of Selection Criteria

The identification of the most relevant articles was carried out in October 2020. The selection was made based on reading article summaries or book introductions. The diagram below shows the item selection process (Fig. 3):

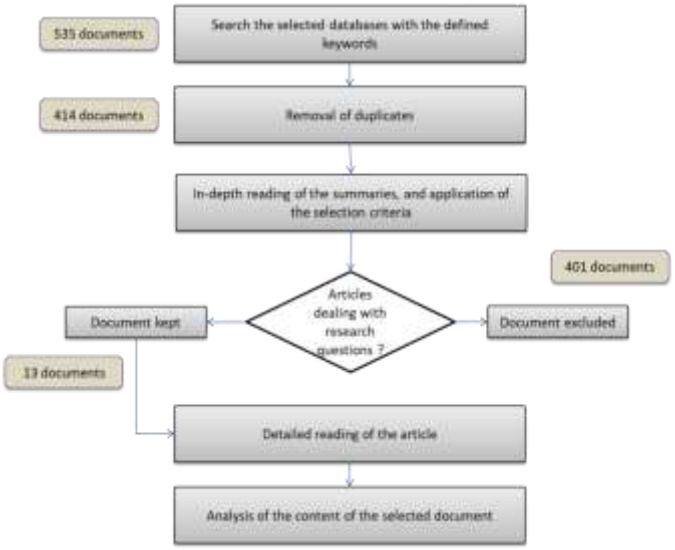


Fig. 3. Research Work Selection Process.

III. Descriptive Analysis of the Results

By analyzing the nature of the research work selected, we find that most of them are journal articles, as the figure below shows (Fig. 4):

On the other hand, the analysis of research work publication's year shows that this is a relatively new subject, since the first article identified was released in 2005. The figure below shows the distribution of the articles, by year of publication (Fig. 5):

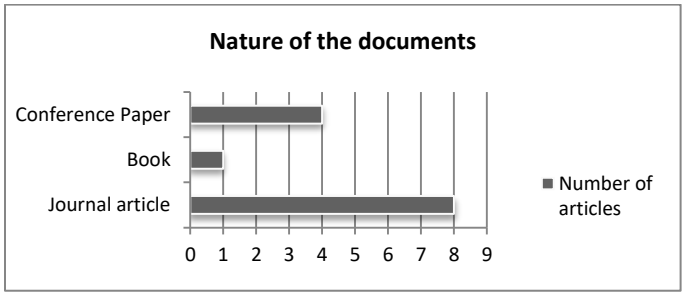


Fig. 4. Number of Documents per Nature.

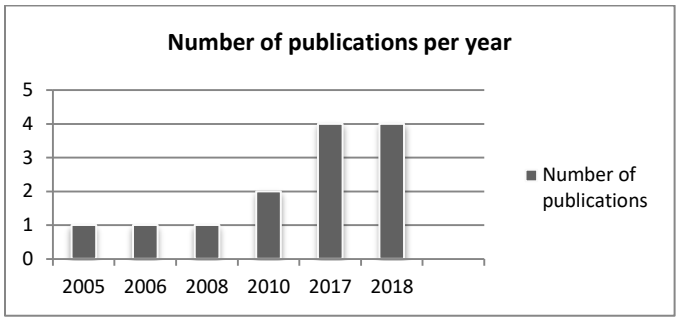


Fig. 5. Distribution of Publications per Year.

Finally, the analysis of the research work type reveals the following results (Fig. 6):

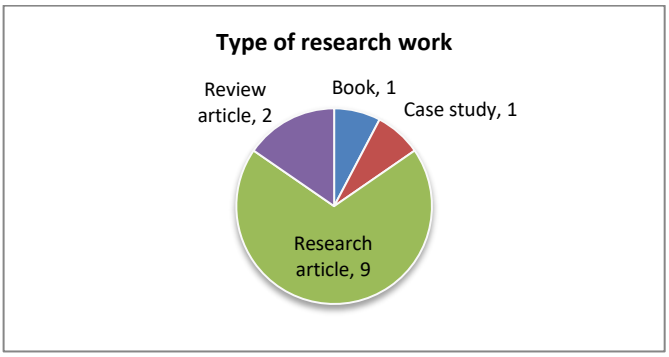


Fig. 6. Distribution of Research Work by Type.

IV. Content Analysis Results

A. Digitizing

Looking at the existing literature, several definitions can be associated with digitization. For some authors, digitalization is the integration of digital technologies into business processes [7][8]. For others, it is an organizational strategy deployed using digital resources to create value [9]. A third definition that seems adequate is that digitizing represents a fundamental change in traditional business practices by redefining new capabilities, processes, and relationships [10].

Despite the diversity of existing definitions, these converge towards the fact that digitization is not limited to the use of digital technologies, but also involves a break with traditional business processes and relationships based on capabilities digital, thus transforming the business model, operational processes and customer experience.

B. Digital Technologies

Several works have tried to list and identify the digital technologies that constitute the basic foundation of digitization. We propose to limit ourselves to the framework proposed by Schlüter and Hetterscheid in 2017, which is oriented towards the digital technologies used at the level of supply chains, and which are as follows [11] (Fig. 7):

Fig. 7. Digital Technologies.

- Mobile devices: A mobile device is a portable tablet or other device designed for portability, and therefore is both compact and lightweight. It's powerful enough to do the same things as a computer.
- Internet of Things (IOT): It is a dynamic global network infrastructure with self-configuration capabilities based on standard and interoperable communication protocols where physical and virtual "objects" have identities, physical attributes and virtual personalities, use intelligent interfaces and are seamlessly integrated into the information network [12]. These infrastructures connect objects and make it possible to access, manage and exploit data. [13].
- Virtual reality is defined as the development of simulated expertise which is somewhat similar to the real-time situation. In addition, virtual reality is an imitation using the computer / communication devices across borders [14]. Augmented reality is a technology that allows you to mix between the virtual world and the real world. It allows digital information to be added to our actual visual field, by superimposition.
- Cloud Computing: A style of computing in which computing capabilities scale and are delivered as a service via internet technologies 'in the cloud'. It also refers to the use of software owned, delivered, and managed remotely by a third party on a paid or subscription basis (also referred to as "software as a service" or "SaaS") [15].
- Blockchain: This is a distributed ledger in which data is stored in a series of blocks. Each of the nodes in a blockchain network has a copy of the blockchain. Each time blockchain data is modified, it is distributed throughout the network [16]. Blockchain technology relies on distributed ledgers, encryption, Merkle tree hashing, and consensus protocols [17].
- Cyber-physical system (CPS): Systems, which directly connect real (physical) objects and processes to (virtual) information processing objects and processes via interconnected, open, and partially global information networks [18].
- Additive manufacturing: It is a process of creating physical objects by superimposing different layers of material on the basis of a digital model. [19].
- Robotics: A robot is a reprogrammable multifunctional manipulator designed to move equipment, parts, tools, or specialized devices by variable programmed movements to perform a variety of tasks. It is a machine capable of automatically performing a complex series of actions programmed by a computer.
- Artificial intelligence: is the means of exploiting, in real time, the mass of information collected, of sorting and analyzing it via algorithms making it possible to build predictive models. This can only be achieved with Big Data, which can be defined as a new generation of technologies designed to enable organizations to extract value from large volumes of data [20].
- Cyber Security: Cyber security corresponds to all the techniques used to preserve the integrity of networks, programs and data against unauthorized access. It refers to all technologies and processes and can also be referred to as information technology security [21].
- Social media: Corresponds to technologies that allow for social interactions, using communication capabilities, such as the internet or a mobile device [15].

C. Impacts of Digital Technologies on the Operational Causes of CFB

1) Demand forecast: One of the major causes behind the CFB is the reliability of demand forecasts and their updating. In fact, forecasts are often based on historical demand. However, distortions in demand can occur, for example, when a customer places an order, the supplier tends to treat this information as a signal of future demand for the product, but in reality the customer's demand often includes a part which is linked to a safety stock that the one always tries to keep in order to protect itself against the variation of demand and delivery times [5], [22]. In this regard, improved demand forecasting and processing of demand signals may be possible through the use of CPS which allows the sharing of operational production information in real time, improving visibility into physical flows [23].

Furthermore, by excluding human interpretation bias, artificial intelligence can improve the processing of demand

signals. Indeed, through the volume of data offered by Big Data, it is easy to develop artificial intelligence to detect and interpret patterns, thus leading to better processing of demand signals [24].

On the other hand, improved processing of demand signals can be achieved using cloud computing. Indeed, by introducing the standards of communication, security and confidentiality, the cloud contributes to the reduction of data inconsistencies that serve as the basis for processing demand signals [25].

That said, another technology also improves the processing of demand signals. Thus, the demand forecasting, which is often carried out by each supply chain member without synchronization, can be significantly improved thanks to the speed of the transmission of information that RFID allows. Indeed, through the data stored at the tag level, RFID makes it possible to provide and process information more quickly and efficiently. This helps to give better visibility in the logistics chain in relation to the flow of materials. This makes integrated planning and therefore better decision making possible [26].

The impact of RFID on improving processing of demand signals has also been highlighted by other researchers. Indeed, the simulation work carried out on a logistics chain, shows that the elimination of the inaccuracy of stocks, which this technology allows, helps to avoid stock-outs [27].

In addition, blockchain is a digital technology that can help improve the processing of demand signals, thanks to the sharing of information it enables between actors in the supply chain. This is because as data sharing is distributed, members of a supply chain can have equal access to data from other members, even when they are further downstream [28].

2) Batch order: Grouping orders in batches can lead to information distortion. Indeed, since placing orders is generally expensive and time consuming, companies prefer to order in batches, instead of ordering their strict needs more frequently [5], [22]. On the other hand, at the production level, manufacturers set the sizes of production batches, to deal with the setup times required to adjust their production machines.

These optimization practices at the local level of each stakeholder in the chain, leads to inefficiency throughout the supply chain thus causing CFB. In this regard, several digital technologies offer solutions to the reduction of batch sizes by reducing the times of operations relating to setup and lunch times.

Thus, the deployment of RFID at the level of the production lines allows the automation of a certain number of operations which are part of the setup times, such as those dedicated to the establishment of tracking sheets, quality sheet, etc. RFID deployed in warehouses contributes to reducing the time and effort required to receive orders, as can be seen from the experience of the Metro distributor, which reduced the time it took to process pallets from 90 to70 seconds through the implementation of RFID [26]. This encourages companies to reduce the size of replenishment order lots.

Another contribution of digital technologies in the reduction of batch sizes is the use of CPS systems for the automation of quality controls, which allows manufacturers to reduce the time dedicated to quality controls which are part of the setup time, and therefore reduce batch sizes [29].

3) Fluctuation in selling prices: Fluctuating prices have an impact on the purchasing decisions of customers. When the price of a product is reduced, customers can buy it in larger quantities than necessary [5], [22]. When the price of the product returns to normal, customers stop buying it until their stocks are depleted. Due to these price fluctuations, the customer's purchasing habits do not reflect an actual consumption pattern, and the variation in quantities purchased is much greater than that in the rate of consumption [5], [22].

This situation, which creates a desynchronization between the rhythm of the order and the rate of consumption of the products, thus gives the supplier an erroneous sign as to the real need of the market. This often causes the supplier to restock (produce) more to keep up with the momentary rise in demand. At the end of the promotions, customers return to their normal rate of demand, thus causing overstock at the level of their supplier, and thus causing the bullwhip effect and consequently the CFB.

Several studies confirm the positive role that certain digital technologies can play in controlling this operational cause of CFB. Thus, the results of the use of algorithms for forecasting demand in a context of variations in selling prices, allow to conclude on the positive impact that artificial intelligence can have in controlling price variations [24].

In addition, the blockchain also makes it possible to store contractual data that the parties can use to detect price fluctuations, in order to correlate them with fluctuations in demand, which contributes to a better analysis of customer demand [28].

4) Rationing games: The rationing games led by members of a supply chain, and which are at the origin of CFB, can be avoided by the use of several digital technologies. Indeed, in the event of a risk of shortage in the market, some companies increase their orders from their suppliers, thus giving a false image on the level of real demand on the market, and pushing suppliers to increase their level of production and stock. To remedy this situation, the deployment of blockchain technology will give suppliers more visibility relative to the actual level of demand, which will allow them to synchronize their production with respect to final demand, rather than blindly following the rationing games of its customers.

5) Lead times: Several research studies highlight the link between the reduction of lead times and the use of digital technologies. Thus, CPS can increase the visibility of physical flows in supply chains and thus improve the availability of information and its sharing. This makes instantaneous data collection and processing possible, reducing production lead times [23], [30]. Additionally, CPS helps in reducing production times by putting feedback loops on production quality [29].

In addition, the speed of data enabled by big data and artificial intelligence also contributes to the reduction of lead times [31].

In addition, the reduction in lead times can be met through the use of cloud computing. Indeed, the sharing of information in real time between the actors of the supply chain, as well as the reduction of data inconsistencies can be ensured by cloud computing [31], [32]. In this sense, the flexibility and scalability of cloud systems provide the necessary infrastructure for simplified information sharing within supply chains, thereby reducing lead times [25].

V. Conclusion

The results of this research allow to conclude that the use of digital technologies can really impact the operational causes of CFB, and lead to its control. Having said that, and looking at the existing research; it seems that only blockchain, artificial intelligence, cloud computing, cyber-physical systems, IOT, and RFID have a role in this control. On the other hand, we found it difficult to link certain technologies such as additive manufacturing, cyber security, augmented reality, and social media with the reduction of CFB. It would be interesting to empirically prove the positive contribution of the technologies previously mentioned on CFB, and more generally on the overall performance of logistics chains. The diagram below gives a synthetic view of the impact of technologies on CFB (Fig. 8):

		Digital technologies				
		CPS	IOT/RFID	Blockchain	Cloud Computing	Artificial Intelligence
Cash Flow Bullwhip Operational Causes	Demand forecast	Instant sharing of production information	Rapid data transfer Data reliability	Collaborative data sharing	Reduction of inconsistencies in demand data via the standards used	Better processing of demand signals via algorithm
	Batch order	Automation of quality controls	Automation of operations at the origin of setup costs			
	Price fluctuations			Management of contractual price data		Adapted algorithms to variations in selling prices
	Rationing games			Real vision of end customer demand		
	Lead times	Automation of quality controls Instant data collection			Real-time data sharing	Rapid data processing through big data

Fig. 8. How Digital Technologies Contribute in the Control of CFB.

References

[1] H. Drissi and M. El ghazali, Audit intene et management des risques, Auditeur. 2018.

[2] M. Fri, F. Fedouaki, K. Douaioui, C. Mabrouki, and E. Semma, "2019, Supply Chain Performance Evaluation Models, State-of-the-Art and Future Directions," Oct. 2019, doi: 10.35940/ijeat.A2049.109119.

[3] A. Gacim, H. Drissi, and A. Namir, "Evaluation of the performance of the university information systems: Case of Moroccan universities," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 224–230, 2019, doi: 10.14569/ijacsa.2019.0100631.

[4] R. Tangsucheeva and V. Prabhu, "Modeling and analysis of cash-flow bullwhip in supply chain," Int. J. Prod. Econ., vol. 145, no. 1, pp. 431–447, 2013, doi: 10.1016/j.ijpe.2013.04.054.

[5] H. L. Lee, V. Padmanabhan, and S. Whang, "The Bullwhip Effect in Supply Chains," Sloan Manage. Rev., vol. Vol. 38, no. 3, 1997.

[6] H. Lamzaouek, H. Drissi, and N. El Haoud, "Cash Flow Bullwhip—Literature Review and Research Perspectives," Logistics, vol. 5, no. 1, p. 8, 2021, doi: 10.3390/logistics5010008.

[7] D. Y. Liu, S. W. Chen, and T. C. Chou, "Resource fit in digital transformation: Lessons learned from the CBC Bank global e-banking project," Manag. Decis., vol. 49, no. 10, pp. 1728–1742, 2011, doi: 10.1108/00251741111183852.

[8] K. Douaioui, M. Fri, C. Mabroukki, and E. A. Semma, "The interaction between industry 4.0 and smart logistics: Concepts and perspectives," 2018 Int. Colloq. Logist. Supply Chain Manag. LOGISTIQUA 2018, vol. 0021266798, no. April, pp. 128–132, 2018, doi: 10.1109/LOGISTIQUA.2018.8428300.

[9] A. B. and O. A. S. and P. A. P. and N. Venkatraman, "Digital business strategy: toward a next generation of insights," Manag. Inf. Syst. Q., vol. 37, pp. 471–482, 2013, doi: 10.1615/TelecomRadEng.v76.i10.20.

[10] H. Lucas, R. Agarwal, E. Clemons, O. Sawy, and B. Weber, "Impactful Research on Transformational Information Technology: An Opportunity to Inform New Audiences," MIS Q., vol. 37, pp. 371–382, Jun. 2013, doi: 10.25300/MISQ/2013/37.2.03.

[11] F. Schlüter, "Supply Chain Process Oriented Technology-Framework for Industry 4 . 0," no. October, 2017.

[12] R. van Kranenburg and S. Dodson, The Internet of Things: A Critique of Ambient Technology and the All-seeing Network of RFID. Institute of Network Cultures, 2008.

[13] B. Dorsemaine, J. P. Gaulier, J. P. Wary, N. Kheir, and P. Urien, "Internet of Things: A Definition and Taxonomy," Proc. - NGMAST 2015 9th Int. Conf. Next Gener. Mob. Appl. Serv. Technol., no. September, pp. 72–77, 2016, doi: 10.1109/NGMAST.2015.71.

[14] M. Javaid and A. Haleem, "Virtual reality applications toward medical field," Clin. Epidemiol. Glob. Heal., vol. 8, no. 2, pp. 600–605, 2020, doi: 10.1016/j.cegh.2019.12.010.

[15] K. Schwertner, "Digital transformation of business," Trakia J. Sci., vol. 15, no. Suppl.1, pp. 388–393, 2017, doi: 10.15547/tjs.2017.s.01.065.

[16] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Cryptogr. Mail. List https//metzdowd.com, Mar. 2009.

[17] P. Tasca and C. J. Tessone, "A Taxonomy of Blockchain Technologies: Principles of Identification and Classification," Ledger, vol. 4, pp. 1–39, 2019, doi: 10.5195/ledger.2019.140.

[18] W. Schroeder, "Germany's Industry 4.0 strategy: Rhine capitalism in the age of digitalisation," Friedrich Ebert Stift., pp. 0–16, 2016, [Online]. Available: https://www.uni-kassel.de/fb05/fileadmin/datas/fb05/FGPolitikwissenschaften/PSBRD/FES-London_Schroeder_Germanys_Industrie_4.0_Strategy.pdf.

[19] "Fabrication additive – connaissez-vous la norme française spécifique à cette technologie industrielle innovante ? - AFNOR Normalisation." https://normalisation.afnor.org/actualites/fabrication-additive-connaissez-vous-la-norme-francaise-specifique-a-cette-technologie-industrielle-innovante/ (accessed Nov. 07, 2020).

[20] R. L. V. C. W. Olofson and M. Eastwood, "Big Data: What It Is and Why You Should Care," International Data Corporation (IDC). .

[21] S. P.S, N. S, and S. M, "Overview of Cyber Security," Ijarcce, vol. 7, no. 11, pp. 125–128, 2018, doi: 10.17148/ijarcce.2018.71127.

[22] Jay W. Forrester, Industrial Dynamics. MIT Press, Cambridge, Mass, 1961.

[23] S. J. Wang, C. T. Huang, W. L. Wang, and Y. H. Chen, "Incorporating ARIMA forecasting and service-level based replenishment in RFID-enabled supply chain," Int. J. Prod. Res., vol. 48, no. 9, pp. 2655–2677, 2010, doi: 10.1080/00207540903564983.

[24] T. O'Donnell, L. Maguire, R. McIvor, and P. Humphreys, "Minimizing the bullwhip effect in a supply chain using genetic algorithms," Int. J. Prod. Res., vol. 44, no. 8, pp. 1523–1543, 2006, doi: 10.1080/00207540500431347.

[25] Y. Yu, R. Q. Cao, and D. Schniederjans, "Cloud computing and its impact on service level: a multi-agent simulation model," Int. J. Prod. Res., vol. 55, no. 15, pp. 4341–4353, 2017, doi: 10.1080/00207543.2016.1251624.

[26] A. Melski, J. Mueller, A. Zeier, and M. Schumann, "Assessing the effects of enhanced supply chain visibility through RFID," 14th Am. Conf. Inf. Syst. AMCIS 2008, vol. 1, no. January 2014, pp. 470–481, 2008.

[27] E. Fleisch and C. Tellkamp, "Inventory inaccuracy and supply chain performance: A simulation study of a retail supply chain," Int. J. Prod. Econ., vol. 95, no. 3, pp. 373–385, 2005, doi: 10.1016/j.ijpe.2004.02.003.

[28] V. Engelenburg, "Delft University of Technology A Blockchain Architecture for Reducing the Bullwhip Effect A blockchain architecture for reducing the bullwhip effect," 2018, doi: 10.1007/978-3-319-94214-8.

[29] L. Baur and E. M. Frazzon, "Evaluating the contribution of in-line metrology to mitigate bullwhip effect in internal supply chains," IFAC-PapersOnLine, vol. 51, no. 11, pp. 1714–1719, 2018, doi: 10.1016/j.ifacol.2018.08.209.

[30] Y. Zhang, L. Zhao, and C. Qian, "Modeling of an IoT-enabled supply chain for perishable food with two-echelon supply hubs," Ind. Manag. Data Syst., vol. 117, no. 9, pp. 1890–1905, 2017, doi: 10.1108/IMDS-10-2016-0456.

[31] T. Eggenberger, K. Oettmeier, and E. Hofmann, "Industrializing Additive Manufacturing - Proceedings of Additive Manufacturing in Products and Applications - AMPA2017," Ind. Addit. Manuf. - Proc. Addit. Manuf. Prod. Appl. - AMPA2017, 2018, doi: 10.1007/978-3-319-66866-6.

[32] C. Gonul Kochan, D. R. Nowicki, B. Sauser, and W. S. Randall, "Impact of cloud-based information sharing on hospital supply chain performance: A system dynamics framework," Int. J. Prod. Econ., vol. 195, pp. 168–185, 2018, doi: 10.1016/j.ijpe.2017.10.008.

IoT System for Vital Signs Monitoring in Suspicious Cases of Covid-19

John Amachi-Choqque[1]
Facultad de Ingeniería y Arquitectura
Universidad Autónoma del Perú, Lima, Perú

Michael Cabanillas-Carbonell[2]
Facultad de Ingeniería
Universidad Privada del Norte, Lima, Perú

***Abstract*—Currently the world is going through a pandemic caused by Covid-19, the World Health Organization recommends to stay isolated from the rest of the people. This research shows the development of a prototype based on the internet of things, which aims to measure three very important aspects: heart rate, blood oxygen saturation and body temperature, these will be measured through sensors that will be connected to a NodeMCU module that integrates a Wi-Fi module, which will transmit the data to an IoT platform through which the data can be displayed, achieving real-time monitoring of the vital signs of the patient suspected of Covid-19.**

***Keywords*—Covid-19; vital signs; internet of things; NodeMCU; IoT platform**

I. Introduction

Currently, the outbreak of the new coronavirus Covid-19, the first case of which was seen in the city of Wuhan, capital of Hubei province (China) [1]. By the end of 2019, it has become a public health problem for the entire globe, since according to data provided by the World Health Organization (WHO), the pandemic is currently present in more than 224 countries, with more than 99,638,000 positive cases and more than 2,140,000 confirmed deaths. The author in [2], increasing by leaps and bounds, thus setting negative records worldwide, due to its high rate of contagion.

The health system in Peru is currently going through a very serious problem, according to the latest reports, there are more than 1,102,000 positive cases and more than 39,800 deaths due to Covid, with a 3.62% lethality rate, with only 11,200 hospitalized patients, 1,892 ICU beds, of which only 7 ventilators are available nationwide. [3], Hospitals and health centers do not have the resources to attend all suspected cases and positive patients. The most recent study on human resources in the health sector indicates that in Peru there are 13.6 physicians for every 10,000 inhabitants, i.e. only 1 physician for 1,000 patients, in addition to an inadequate distribution of medical personnel at the national level. [4], making the healthcare system totally deficient and inadequate to deal with the increasing number of patients caused by Covid-19.

As a result of the aforementioned data, the following question arises: What happens to the people who tested positive for Covid-19, because although Covid-19 cases are classified into five stages: asymptomatic, mild, moderate, severe and critical? [5]. It is those in serious and critical condition that are treated in health centers. Once the patient has been diagnosed with Covid-19, he/she is obliged to remain isolated in his/her home until the incubation and infection stage has passed, which can last between 12 to 15 days. [6], in addition to maintaining distance from family members to reduce the likelihood of contagion.

A new question arises: What happens to patients who are isolated in their homes, because they suffer the risk that the disease caused by Covid-19 worsens, and if they are not administered the necessary drugs, they may die, to perform this follow-up they would normally have to be taken to the hospital, where they will undergo various tests to identify the heart rate, respiratory rate, blood oxygen saturation, blood pressure and body temperature, because Covid-19 to develop in the body [7], the health system carries out patient follow-ups by medical personnel, who go to the homes of positive or suspected Covid cases, where the lives of medical personnel are exposed to contracting the disease, in addition to generating effort and expense in the process.

Given the current situation in Peru, many of the medical centers nationwide are full of patients, exceeding their capacity of care, in these circumstances the medical centers do not attend in the right way, so people have to opt for private health services, as is the case of clinics, However, low-income people cannot have access to this service, neither to health services, nor to a Covid screening, having to spend the incubation stage in their homes, keeping home isolation, increasing the number of people vulnerable to contracting Covid-19, exposing the family of the infected or suspected case, if the necessary measures are not taken such as: isolating the infected person, keeping a distance of at least 2 meters and controlling symptoms on a daily basis. Faced with this situation that the country is going through, it is necessary to resort to innovative and outstanding ideas for the solution of the different problems that this pandemic has generated in society.

As we know, the internet of things has been developed even in the health sector, called telemedicine [8], but it can also be applied in the same homes for medical and health purposes. That is why an internet of things system will be developed to monitor vital signs in patients or suspected cases of Covid-19, this is done with the help of different specialized sensors.

The internet of things (IoT) is the interconnection of devices (sensors and actuators) or objects (everyday objects with internet access) through a network, in order to

communicate and transfer information, without the need for human presence to do so, this is called machine-machine communication (M2M), for the development of an IoT system protocols, communication technologies, domains and applications are established [9]. The proposed IoT system aims to measure certain vital signs in order to provide prompt help in case of any drastic change in their health, reducing the effort of medical staff [10], also avoiding that the patient goes through stress, produced when a person is hospitalized, in the same way, reducing stress in medical personnel, according to a study done in China, cases of 1257 workers are reported, 50% began to feel symptoms of depression and more than 70% presented symptoms of psychological distress [11] thus generating a high risk for those who face this pandemic in the first row, with this proposed solution, the time to obtain vital signs, the time of medical care of home visits and the response time to an anomaly in the vital signs are reduced.

II. Theoretical Framework

A. Internet of Things (IoT)

Also known as IoT. It is the interconnection between devices, objects or things-electrical appliances, modules, machines, devices and more, through the internet to communicate and exchange data [12].

To make it possible to develop this technology it is necessary an integral series of technologies, such as the (API) that are those that connect to the internet the different devices, in addition to the use of standards and IoT platforms where the different devices that are connected will be visualized.

B. Arduino IDE

It is an open source Arduino software, where it is easier to code, load and run a series of codes, which will form the program we develop, this software is a text editor and compiler at the same time, serves to program and to transfer the code to the Arduino board, but is also compatible with many other modules, note that this software works with the Processing programming language and can be installed on operating systems such as Windows, Mac and Linux [13].

C. NodeMCU ESP8266

It is a development board belonging to the NodeMCU family, it is a totally free software and hardware, this board allows the connection of several devices with each other, through the internet, thanks to the ESP8266 Wi-Fi module that has incorporated, this chip is also compatible with TCP/IP, being the easiest and fastest way to develop IoT projects [14].

D. Pulse Oximeter Sensor MAX30102

Very compact sensor, it is considered non-invasive, with which you can measure: the level of oxygen saturation in hemoglobin (SpO2), through a LED circuit and a photodetector capable of measuring the amount of light reflected through the finger, as there are variations between the reflection that occurs through the blood loaded with oxygen with deoxygenated blood, oxygenated blood tends to absorb more infrared light, while deoxygenated blood absorbs more red light [15].

E. Sensor LM35

It is a temperature sensor of good assertiveness index, having a very low cost, with a working range between -55°C to 150°C, has an analog output with its respective power pins, has an accuracy of 0.5°C making it easy to use with a variety of applications where it can be implemented [16].

III. Background

In recent years, medicine has made great advances, developing a variety of technologies for health monitoring. At [17] the importance of the development of portable biomedical sensors to facilitate the remote monitoring of patients is expressed, focusing on the measurement of heart rate and body temperature, for different conditions presented by the patient, whose data will be sent to a doctor through the Zigbee network.

In the investigation [18] presents a monitoring system developed for the measurement of cardiac pulse and oxygen saturation, focused on preventing and monitoring different diseases, consisting of an oximetry sensor and a Nellcor DS-100 sensor in charge of detecting the signs and their variations, which will then be sent to a mobile application, which will process the data to issue an alarm if necessary and to visualize the data.

In the investigation [19] shows a prototype system for monitoring vital signs, including body temperature, heart rate and oxygen desaturation, using an lm35 sensor, Pulse Sensor Amped and an Arduino board, which will allow the detection, processing and sending data to a mobile application on a cell phone and with a monitor you can view the graphs with respect to the heart rate.

In the investigation [20] presents a system to monitor the patient's desaturation remotely, taking into account the anomalies that can cause oxygen desaturation in a patient, so it is considered a permanent monitoring of this sign, to improve the diagnostic process of the patient, who is at home, this is achieved through different electronic components and a Wi-Fi module, at the end the data are displayed on a local host by the doctor.

In all the mentioned works is present the importance of monitoring vital signs, focused on different types of diseases, the research work developed, focuses on the monitoring of Covid-19, establishing the different levels of severity of the different signs, has the minimum number of sensors to capture the signs that vary in that disease, In addition to having a web system, in which the doctor can view statistics and graphs of the different vital signs, managing to send alerts to an email or a mobile device, in this way, this research unifies different technological aspects of the antecedents and seeks better alternatives to focus on monitoring patients suspected of Covid-19.

IV. Methodology

For the development of this project, Methodology V is used, this methodology is used for the development of ICT projects, used for the management and also for the development of systems, especially software development for ICT components.

The reason why this methodology is used is because it is very easy to use for the development of this research, has 6 phases, focuses on quality management procedures, because at each level it has, there is an opposite side that performs the tests and thus reduce the risks that the project or product goes wrong.

The V methodology or V method, has 4 levels of which there is a parallel phase of verification, referring to the shape of the model, as it compares the phases of development with their respective quality control, in each phase describes the activities performed and the results produced throughout the development, on the left side are the phases of specification, containing the tasks of design and development of the system, while on the right side are the phases of testing, which contain the control measures of each phase as unit tests and integration tests [21].

V. Development

The IoT system that was developed, can perform three measurements of a patient, which are: heart rate, oxygen saturation and body temperature, these being the vital signs that are affected by the disease Covid-19, presenting with various symptoms such as fever, cough among others, the data of these vital signs, are obtained through the use of two sensors such as the MAX30102 sensor, with which it is possible to obtain the heart rate and oxygen saturation in the blood and the LM35 sensor for measuring body temperature.

A. Phase 1: Specifications

At this stage the appropriate sensors and modules will be chosen, among the sensors is the MAX30102, which is used to measure heart rate and oxygen saturation, through red and infrared LED circuit, both lights are intercepted are reflected through the finger and a photodiode captures it and calculations are made of oxygen saturation, can also measure heart rate while held down, after a series of operations are performed, such as calculating the average per minute, you get a specific number, this being the number of beats per minute [22]. The other sensor to be used is the LM35 sensor that can measure the temperature, with a high assertiveness, having an accuracy of 0.5°C.

We also use the NodeMCU board, which has a microcontroller, where it is programmed through the Arduino IDE. The important thing about this board is that it has a WiFi module, this allows a wireless connection that can easily connect us to the internet, allowing us to send or receive files over the internet.

Table I shows the functional requirements of the prototype, then Table II shows the non-functional requirements of the prototype to be developed.

B. Phase 2: Overall Design

The design of the system is presented interacting with the web and those involved, visualizing the path of the vital signs data, it starts when the patient is using the prototype and this is connected to the WiFi network, where the NodeMCU board will be the brain of the system, thanks to the microcontroller it has, in this module will be connected LM35 sensors for temperature and MAX30102 sensor for heart rate and oxygen saturation, then the data will go to the internet where they reach the Ubidots platform and then to the web application.

All the above process is reflected in the system architecture, as shown in Fig. 1.

C. Phase 3: Detailed Design

1) Prototype design: Fig. 2 shows the established design, using the Fritzing software, the connections between the NodeMCU board and the MAX30102 and LM35 sensors can be observed.

TABLE I. Prototype Functional Requirements of the Prototype

Functional Requirements	
RFP1	The prototype must connect to the home WiFi network automatically.
RFP2	The prototype must be connected to the Ubidots platform.
RFP3	The prototype must analyze serial communication data from the MAX30102 sensor to obtain frequency and saturation data.
RFP4	The prototype must analyze the serial communication data from the LM35 sensor to obtain the temperature data.

TABLE II. Non-Functional Requirements of the Prototype

Non-functional requirements	
RNFP1	The prototype must be connected to the MAX30102 sensor.
RNFP2	The prototype must be connected to the LM35 sensor.
RNFP3	The prototype must integrate a NodeMCU
RNFP4	Prototype must be connected to a battery for portability.
RNFP5	The prototype must not be invasive to the user.
RNFP6	The prototype should start operating when connected to a power source.

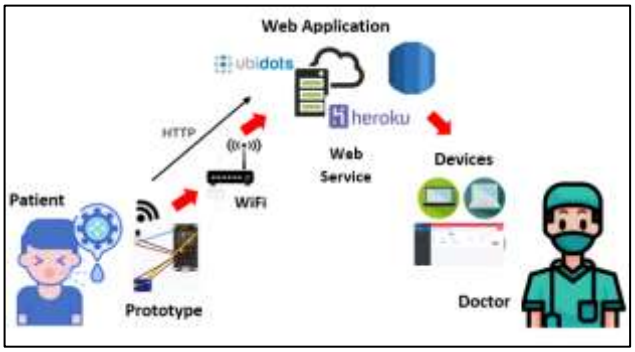


Fig. 1. System Architecture.

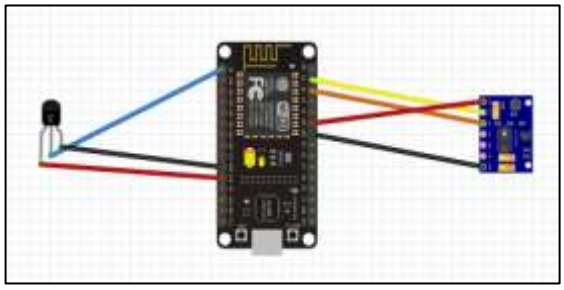


Fig. 2. Connection of the Prototype in Fritzing.

D. Phase 4: Implementation

1) LM35 sensor conditioning: The LM35 temperature sensor is very small and has three pins for connection. When connected with a typical jumper cable, it does not connect firmly to the pins of the LM35 and when it comes in contact with the skin, it alters the captured data.

It is therefore necessary to solder the pins and cover them with thermofit, as shown in Fig. 3, where the LM35 sensor pins are completely covered.

2) LM35 sensor programming: The LM35 sensor is programmed to obtain the body temperature data, in addition to establishing the Wi-Fi connection, the HTTP protocol must be established to send the data to the Ubidots platform, a certain part of the code is as shown in the following Fig. 4.

3) MAX30102 sensor programming: The MAX30102 sensor is programmed to obtain the data related to the oxygen saturation and heart rate signals. Fig. 5 shows part of the code, specifically the Void Loop, where the data that will be printed on the serial monitor at the time of execution can be seen.

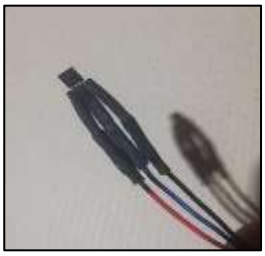


Fig. 3. Conditioned LM35 Sensor.

```
#include "Ubidots.h"

const char* UBIDOTS_TOKEN = "[redacted]";  // Put here your
const char* WIFI_SSID = "MOVISTAR_1618";      // Put here your Wi-Fi SSID
const char* WIFI_PASS = "[redacted]";      // Put here your Wi-Fi password
const char* DEVICE_LABEL ="NodeMCU";
Ubidots ubidots(UBIDOTS_TOKEN, UBI_HTTP);


float TEMPERATURA;
float SUMA;
//////////////////////////////////

void setup() {
  Serial.begin(115200);
  ubidots.wifiConnect(WIFI_SSID, WIFI_PASS);
  // ubidots.setDebug(true); // Uncomment this line for printing debug messages
}
```

Fig. 4. LM35 Sensor Programming.

```
//send samples and calculation result to terminal program through UART
Serial.print(F("red="));
Serial.print(redBuffer[i], DEC);
Serial.print(F(", ir="));
Serial.print(irBuffer[i], DEC);

Serial.print(F(", HR="));
Serial.print(heartRate, DEC);

Serial.print(F(", HRvalid="));
Serial.print(validHeartRate, DEC);

Serial.print(F(", SPO2="));
Serial.print(spo2, DEC);

Serial.print(F(", SPO2Valid="));
Serial.println(validSPO2, DEC);
```

Fig. 5. MAX30102 Sensor Programming.

E. Phase 5: Verification

This phase focuses on the verification of the hardware and software modules, on a unitary basis, to check that they are working properly.

1) MAX30102 sensor test: After programming the MAX30102 sensor, the necessary test is performed, where the oxygen saturation data captured by the sensor is observed through the serial monitor, as shown in Fig. 6.

2) LM35 sensor test: After programming the LM35 sensor, we proceed to perform the necessary tests. In Fig. 7, the patient data captured by the sensor are shown, printed on the serial monitor, where the values must be stable, so that the values are automatically sent to the Ubidots platform.

F. Phase 6: Integration

In this phase, the sensors will be integrated into the same code in the Arduino IDE, and the data will be obtained at the same time from the NodeMCU board, so that they are then sent to the Ubidots platform and subsequently sent through the web system, thus, to perform the respective monitoring of the patient.

1) Connection of the prototype: We proceed to connect all the system components: the NodeMCU board, the MAX30102 sensor and the LM35 temperature sensor. Fig. 8 shows the prototype with all the components correctly connected.

2) Unified prototype programming: After making the prototype connections, integrating the NodeMCU board with the MAX30102 and LM35 sensors, we proceed to perform the necessary programming.

```
red=9562, ir=65445, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9613, ir=65438, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9637, ir=65327, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9595, ir=65326, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9483, ir=65150, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9202, ir=65141, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9531, ir=65302, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9552, ir=65319, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9570, ir=65170, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9507, ir=65245, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
red=9632, ir=65233, HR=75, HRvalid=1, SPO2=97, SPO2Valid=1
```

Fig. 6. Sensor Test Max30102.

```
.....WiFi connected
IP address:
192.168.1.10
26.9
27.1
27.3
28.9
31.3
32.5
33.6
34.4
Valor enviado a Ubidots
35.0
Valor enviado a Ubidots
35.7
Valor enviado a Ubidots
35.8
Valor enviado a Ubidots
35.7
```

Fig. 7. Testing the LM35 Sensor.

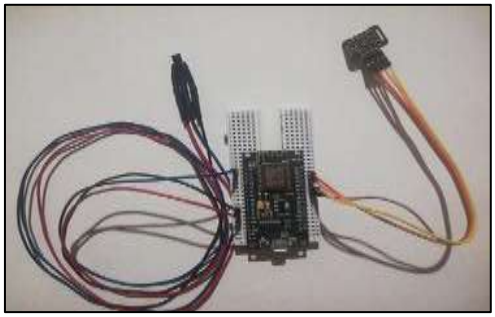


Fig. 8. Connection of the IoT Prototype.

As shown in Fig. 9, we can see the libraries and variables established to operate both sensors, in addition to placing the credentials to establish the Wi-Fi connection and access to the Internet.

Fig. 10 shows some operations established in the Void Loop, necessary for sending data to the Ubidots platform, the data are captured and sent in a specific order, in addition to containing an identifier for that prototype.

3) Prototype testing: The test is performed, where the data must be real values of the patient, and the test is performed by comparing the printed values with the values obtained by means of specialized instruments.

Fig. 11 shows the data printed on the serial monitor, the data are captured by the sensors and sent to the Ubidots platform, in addition to having established a condition not to capture data if the patient does not have his finger on the MAX30102 sensor, thus avoiding the sending of invalid data.

As shown in Fig. 12, the Ubidots platform can be seen, with the data sent by the prototype.

The data from the Ubidots platform are sent to the developed web system, where the patient chosen to perform the test is selected, and the respective values of the patient's vital signs are displayed, as shown in Fig. 13.

The patient data in the web system can also be viewed through graphs, as shown in Fig. 14.

```
#include <Wire.h>
#include "MAX30105.h"
#include "Ubidots.h"
#include "heartRate.h"
#include "spo2_algorithm.h"

MAX30105 particleSensor;

const char* UBIDOTS_TOKEN = "[redacted]";
const char* WIFI_SSID = "MOVISTAR_1610";
const char* WIFI_PASS = "8bzaGjfpuJnOeHgd7Nry";
const char* DEVICE_LABEL ="NodeMCU";
Ubidots ubidots(UBIDOTS_TOKEN, UBI_HTTP);

uint16_t irBuffer[100];
uint16_t redBuffer[100];
#else
uint32_t irBuffer[100];
uint32_t redBuffer[100];
#endif
```

Fig. 9. Libraries and Variables.

```
if((bpm>40)&&(temp>34)&&(spo2>80))
{
    ubidots.addContext("CODE_UNICO","9234e17c-3700-11eb-adc1-0242ac120002");
char* context = (char*)malloc(sizeof(char) * 40);
 ubidots.getContext(context);

 ubidots.add("saturacion", spo2, context);
 ubidots.add("frecuencia", bpm,context);
 ubidots.add("temperature", temp,context);

 bool bufferSent = false;
 bufferSent = ubidots.send(DEVICE_LABEL);
   if (bufferSent) {
  Serial.println("Valor enviado a Ubidots");
   }
free(context);
}
}
```

Fig. 10. Code for Sending Data to Ubidots.

```
Valor enviado a Ubidots
IR=115328, SpO2=102, BPM=67, TC=35.89

Valor enviado a Ubidots
IR=115071, SpO2=102, BPM=67, TC=36.14

Valor enviado a Ubidots
IR=115036, SpO2=102, BPM=67, TC=36.40

Valor enviado a Ubidots
IR=114816, SpO2=102, BPM=67, TC=36.65

Valor enviado a Ubidots
IR=115026, SpO2=102, BPM=67, TC=36.71
```

Fig. 11. Values Sent to Ubidots.

Fig. 12. Display of Values Sent to Ubidots.

Fecha	Hora	SpO2	TC	BPM
14/12/2020	17:08:06	97	36.65	62
14/12/2020	17:08:20	97	36.97	62
14/12/2020	17:08:40	97	36.97	62
14/12/2020	17:08:47	97	36.97	62
14/12/2020	17:09:24	97	37.28	62

Fig. 13. Data Displayed in the Web System.

Fig. 14. Graphing of Data in the Web System.

VI. Results

Different tests were performed to test the hypothesis: that the use of a prototype system based on the Internet of Things improves the process of monitoring vital signs in suspected cases of Covid-19, taking into account the following indicators:

Indicator 1: The use of a prototype system based on the internet of things reduces the time required to obtain vital signs.

In this test we want to demonstrate that the time in which the measurements of the patient's vital signs are taken, using the patient's daily instruments and having to go to the patient's home to take the measurements, takes an average of 29.5 minutes, while using the prototype developed and taking the measurements remotely, the average is 4.6 minutes, as shown in Fig. 15.

Indicator 2: The use of a prototype system based on the Internet of Things reduces the medical care time.

In this test we want to show that the time in which the patient receives medical care, including results, recommendations and prescription, is expedited, this process takes on average about 28.3 minutes, instead using the prototype developed and perform medical care remotely, the average is 12.4 minutes, as shown in Fig. 16.

Indicator 3: The use of a prototype system based on the internet of things reduces the communication time in the event of an anomaly in the vital signs.

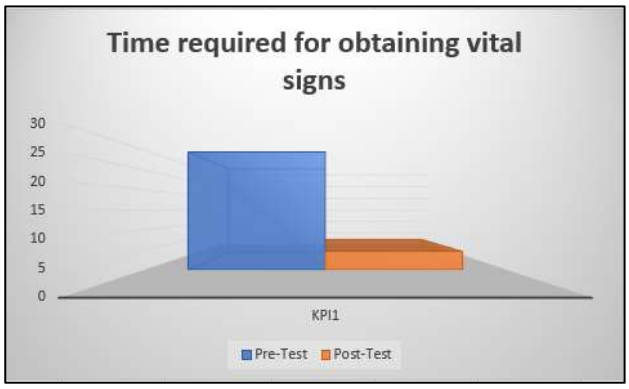


Fig. 15. KPI1 Pre-Test and Post-Test Comparison.

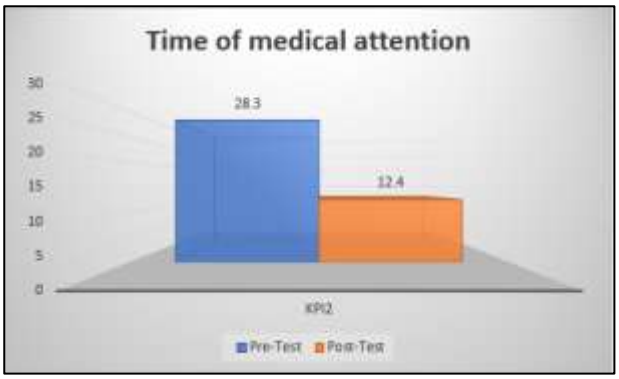


Fig. 16. KPI2 Pre-Test and Post-Test Comparison.

In this test it is desired to show that the time in which the doctor is alerted about any change in the patient's signs, so that the doctor can perform the necessary actions in relation to his condition, this process takes on average about 2528.5 seconds, instead using the developed prototype, which can issue a notification, message or call, the average communication time of the patient's condition is 18.8 seconds, as shown in Fig. 17.

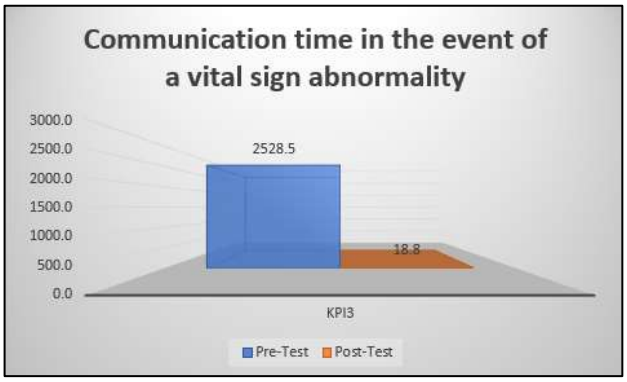


Fig. 17. KPI3 Pre-Test and Post-Test Comparison.

VII. Conclusions

In conclusion, this system is of great help to patients or suspected cases of Covid-19, who are in their homes, monitoring their vital signs and see if the disease worsens in their body to provide a prompt solution and provide appropriate assistance, preventing the disease from worsening in their body.

In a different way, it will help medical personnel by preventing them from being exposed to many Covid-19 positive cases and running the risk of becoming infected, eliminating the face-to-face follow-ups that are currently performed on positive patients who are isolated in their homes. While it is true that medical staff will always be needed in hospitals or clinics, this system will help to reduce the number of patients presenting at hospitals, and staff can focus only on severe cases and reduce to some extent their stress and fear of exposure and interaction with so many patients.

An internet of things prototype, built with the NodeMCU board, is beneficial for monitoring vital signs and facilitates sending data to the internet in a fast and secure way, with the different communication protocols that can be used to send data to the internet.

The construction of a vital signs monitoring system can be done with a minimum amount of sensors and expenses, since the sensors and boards used in this project are the cheapest in the market, and the code for its operation can be found by searching the internet.

References

[1] "New coronavirus 2019." https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019 (accessed May 31, 2020).

[2] "Coronavirus disease 2019." https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed May 31, 2020).

[3] "Covid 19 en el Perú - Ministerio del Salud." https://covid19.minsa.gob.pe/sala_situacional.asp (accessed Jun. 01, 2020).

[4] J. C. Loayza Altamirano, M. L. Chilca Alva, and W. Pérez Lázaro, [Statistical Compendium: Human Resources Information of the Health Sector, Perú 2013 - 2018] author's translation. 2019.

[5] X. R. Ding et al., "Wearable Sensing and Telehealth Technology with Potential Applications in the Coronavirus Pandemic," IEEE Rev. Biomed. Eng., 2020, doi: 10.1109/RBME.2020.2992838.

[6] Ministerio de Sanidad, "Aportaciones de esta actualización INFORMACIÓN CIENTÍFICA-TÉCNICA Enfermedad por coronavirus, COVID-19," 2020. Accessed: Nov. 27, 2020. [Online]. Available: https://www.aemps.gob.es/.

[7] OMS, "Questions and answers on coronavirus disease (COVID-19)," 2020. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses (accessed Nov. 27, 2020).

[8] J. Leng, Z. Lin, and P. Wang, "Poster abstract: An implementation of an internet of things system for smart hospitals," Proc. - 5th ACM/IEEE Conf. Internet Things Des. Implementation, IoTDI 2020, pp. 254–255, 2020, doi: 10.1109/IoTDI49375.2020.00034.

[9] S. DIvakaran, L. Manukonda, N. Sravya, M. M. Morais, and P. Janani, "IOT clinic-Internet based patient monitoring and diagnosis system," IEEE Int. Conf. Power, Control. Signals Instrum. Eng. ICPCSI 2017, pp. 2858–2862, 2018, doi: 10.1109/ICPCSI.2017.8392243.

[10] S. Pradeep Kumar, V. R. R. Samson, U. B. Sai, P. L. S. D. Malleswara Rao, and K. Kedar Eswar, "Smart health monitoring system of patient through IoT," Proc. Int. Conf. IoT Soc. Mobile, Anal. Cloud, I-SMAC 2017, pp. 551–556, 2017, doi: 10.1109/I-SMAC.2017.8058240.

[11] J. Lai et al., "Factors Associated With Mental Health Outcomes Among Health Care Workers Exposed to Coronavirus Disease 2019," JAMA network open, vol. 3, no. 3. p. e203976, 2020, doi: 10.1001/jamanetworkopen.2020.3976.

[12] K. Rose, S. Eldridge, and L. Chapin, "OCTUBRE DE 2015 Para entender mejor los problemas y desafíos de un mundo más conectado," 2015. Accessed: Nov. 27, 2020. [Online]. Available: https://www.internetsociety.org/wp-content/uploads/2017/09/report-InternetOfThings-20160817-es-1.pdf.

[13] "IDE – Aprendiendo Arduino." https://www.aprendiendoarduino.com/tag/ide/ (accessed Jan. 28, 2021).

[14] Naylamp Mechatronics NodeMCU, "NodeMCU v2 ESP8266 WiFi ," 2020. https://www.naylampmechatronics.com/espressif-esp/153-nodemcu-v2-esp8266-wifi.html (accessed Dec. 01, 2020).

[15] Maxim Integrated, "MAX30102," 2018. Accessed: Dec. 01, 2020. [Online]. Available: www.maximintegrated.com.

[16] Naylamp Mechatronics, "Sensor de Temperatura analógico LM35," 2020. https://www.naylampmechatronics.com/sensores-temperatura-y-humedad/234-sensor-de-temperatura-analogico-lm35.html?search_query=lm35&results=4 (accessed Dec. 01, 2020).

[17] S. Sali and C. S. Parvathi, "Integrated wireless instrument for heart rate and body temperature measurement," 2017 2nd Int. Conf. Converg. Technol. I2CT 2017, vol. 2017-Janua, pp. 457–463, 2017, doi: 10.1109/I2CT.2017.8226171.

[18] O. I. Arias Juárez, "Diseño e implementación de un sistema de monitoreo para la medición del pulso cardíaco y saturación de oxígeno en la sangre," Quito: Universidad de las Américas, 2017, 2017.

[19] L. E. Chunga Limo and L. F. Roa Martínez, "Diseño e implementación de un prototipo para un sistema de monitoreo de signos vitales con aplicación para dispositivos móviles," Universidad Nacional Pedro Ruiz Gallo, 2019.

[20] J. Calderón Quispe, "Implementación de un oximetro de pulsos para monitorizar la desaturación del paciente a distancia," 2019.

[21] IONOS España, "Modelo V: definición, ventajas y áreas de aplicación - IONOS," Jun. 2020. https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/modelo-v/ (accessed Dec. 08, 2020).

[22] Kerim Bedri Saçan dan Gökhan Erta, "MAX30100 SpO 2 / Nab Õ z Duyargas Õ n Õ n Performans De ÷ erlendirmesi Performance Assessment of MAX30100 SpO 2 / Heartrate Sensor," 2017.

Factors Influencing Master Data Quality: A Systematic Review

Azira Ibrahim[1], Ibrahim Mohamed[2], Nurhizam Safie Mohd Satar[3]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia

***Abstract*—Master data refers to the data that represents the core business of the organization, shared among different applications, departments, and organizations and most valued as the important asset to the organization. Despite the outward benefit of master data mainly in decision making and organization performance, the quality of master data is at risk. This is due to the critical challenges in managing master data quality the organization may expose. Hence the primary aim of this study is to identify factors influencing master data quality from the lens of total quality management while adopting the systematic literature review method. The study proposed 19 factors that inhibit the quality of master data namely data governance, information system, data quality policy and standard, data quality assessment, integration, continuous improvement, teamwork, data quality vision and strategy, understanding of the systems and data quality, data architecture management, personnel competency, top management support, business driver, legislation, information security management, training, change management, customer focus, and data supplier management that can be categorized to five components which are organizational, managerial, stakeholder, technological, and external. Another important finding is the identification of the differences for factors influencing master data compared to other data domain which are business driver, organizational structure, organizational culture, performance evaluation and rewards, evaluate cost/benefit tradeoffs, physical environment, risk management, storage management, usage of data, internal control, input control, staff participation, middle management's commitment, the role of data quality and data quality manager, audit, and personnel relation. It is expected that the findings of this study will contribute to a deeper understanding of the factors that will lead to an improved master data quality.**

Keywords*—*Quality management; total quality management; data quality; data quality management; master data; master data quality; master data quality management; systematic literature review

I. Introduction

The evolution of digital transformation and a data-driven economy requires the formulation of new strategies to ensure the organization stays relevant and competitive. An organization is expected to face various issues as the effect of development that requires proactive management action [1]. Taking into account that data is an important element for every organization [2]–[4], the massive amount of data that are created and stored in response to digitalization possess new challenges in the management of data quality.

We greatly appreciate funding received from Universiti Kebangsaan Malaysia (ETP-2013-060) and Malaysian Public Service.

In particular, the organization is normally held responsible to manage a few types of data namely master data, transaction data, and reference data, to name a few. Master data is ranked as having the highest priority to be managed due to the valuable information it holds about the organization [5] and should be considered as an important asset to the organization [1], [6]. Master data represents the organization's core business objects that form the foundation of the main business process and must therefore be used unambiguously across the entire related application, department, and organization. Typical master data classes are supplier, customer, material, product, employee, and asset [7]–[9]. In the public sector context, master data composed of data about service providers, customers, and services or products offered [10].

The importance of master data requires it of high quality in supporting the organization to perform roles such as planning and decision making [11] and ensuring compliance with the regulatory and legal provision [12]. While the increasing demand for information system initiatives evidenced that high-quality master data is one of the important elements in the successfulness of the implementation [13], [14]. According to [12] current, accurate, and complete master data is required.

Studies in academic and industry highlighted that data quality is an urgent issue. The impact of poor data quality can be manifested across the operational, tactical, and strategic levels of the organization [15]. In the specific context of master data, poor master data quality incurred additional costs to the organization which involves a cost in assuring the quality of master data and cost affected by poor data quality [16]. On a similar tone, The Data Warehousing Institute (TDWI) calculated that data quality problems cost U.S. businesses about USD 600 billion a year [17]. Similarly, a study conducted in 2016 by Royal Mail [18] showed that poor quality of customer contact data costs, on average, 5.9% of the annual revenue to UK companies.

Despite the benefit and impact of poor master data quality, improving master data quality is still an issue. The industry is struggling in trusting the quality of the data and the implementation of data quality measures. A recent survey evidenced that only 40% of the respondent confident in the quality of data in their company and also their organization's data quality management practices [19]. According to [20] poor master data quality is one of the biggest challenges faced by the organization in managing the complexity of digitalization apart from standardization and governance.

Furthermore, according to [1], 80% of companies acknowledged the impact of poor master data quality to be high or very high for their performance, 82% of the company engaged in data quality initiative but not using the systematic or established method and only 15% of the companies know the established method for improving master data quality.

Undoubtedly, the importance of master data, the effect of poor quality master data, and the lack of adequate master data quality management underline the importance to initiate a study that revolves around the establishment of systematic master data quality management in ensuring the improvement of master data quality. However, considering that master data appears to have different characteristics compared to other domains of data and featuring distinct challenge and requirement, such as organizational, people, process, and technology [7], [21]–[23], thus a deeper understanding of the aspect related to master data is required before commencing any improvement initiative.

Xu [24] highlighted the importance to investigate, understand, and explain the factors influencing data quality, before proceeding with data quality improvement. However, a study that systematically explores factors influencing master data quality is scarce. Fortunately, the progress in the data quality management discipline by [24]–[30] made a substantial contribution in investigating factors influencing data quality.

The theoretical foundation for data quality management studies was originated from the Total Quality Management (TQM) discipline. TQM originally focused on quality improvement in the manufacturing domain [31], [32]. TQM provides an established organizational-wide foundation in identifying factors that contribute to data quality in the organization namely stakeholder, quality management, teamwork, process management, and top management support [33]. Based on TQM, [34] introduced the Total Data Quality Management (TDQM) approach in managing data quality, with the analogy of data as a manufactured product. The contribution by [34], is regarded as an important milestone for the advancement in data quality study.

In response, this paper aims to identify factors influencing master data quality from the lens of TQM based on the current and rigorous work in data quality management. The identification of the factors influencing master data quality will support the ongoing study in developing a framework for managing master data quality. Therefore, yields two research questions which are 1) what are the factors influencing master data quality in the organization?, and 2) how do the factors influencing master data quality differ from other data domains?. This paper employs systematic literature review approach in answering both research questions.

The remainder of the paper is structured as follows: Section II reviews the literature on data quality and master data quality. Section III describes the method for conducting a systematic literature review. Section IV presents the finding of the study. Section V discusses the finding. The paper ends with conclusions in Section VI.

II. Related Work

A. Data Quality

Data quality is a complex construct composed of multiple dimensions [35]–[39]. Although previous scholars agree that there is no definite definition for data quality, however, it was acknowledged that data quality must meet user requirements for specific usage context or fitness for use [40]–[42]. Seminal literature such as [37] operationalized the term data quality using dimensions namely accuracy, timeliness, completeness, and consistency.

While defining data quality is an issue, the same goes for identifying the factors influencing data quality. Grounded on the theory of TQM, the studies in data quality progressively contribute to a deeper understanding of issues related to data quality. Besides, data quality can be considered as a sub-discipline of TQM. Several researchers show the advancement in discussing factors influencing data quality in various contexts [24]–[28]. Based on the theory of TQM, factors influencing data quality can be classified into five components which are organizational, managerial, stakeholder, technological, and external [25], [43], [44].

The works by [24], [25] focusing on the quality of accounting data that resides in AIS were among the most cited work in understanding factors influencing data quality. The theoretical foundations of the study are based on four area which are TQM, just-in-time (JIT), data quality, and accounting.

In getting a deeper insight into the factors influencing accounting data quality, [25] applied a qualitative methodology involving multiple case studies. The author suggested 26 factors that were classified by five categories, namely 1) AIS characteristics (nature of system), 2) data quality characteristics (data quality policies and standards, data quality approach, role of data quality, internal control, input control, understanding of the system and data quality, and continuous improvement), 3) stakeholders (top management's commitment, middle management's commitment, roles of data quality manager/manager group, customer focus, personnel relations, information supplier management, audit and review, and personnel competency), 4) organizational (training, organizational structure, organizational culture, performance evaluation and rewards, manage change, evaluate cost/benefit tradeoffs, teamwork, physical environment, and risk management), and 5) external factor.

Complementing the study by [25], the three most important factors influencing accounting data quality suggested by [24] through quantitative study namely 1) top management commitment, 2) the nature of the systems, and 3) input controls. Further, in the context of health data, [26] suggested six factors influencing data quality which are 1) top management support, 2) resources, 3) regulatory capability, 4) business-IT alignment, 5) staff participation, and 6) data/system integration.

In contrary to the previous studies, [27], [28] explored factors related to data quality management regardless of specific data domain, where the findings support a higher

generalization. Also rooted in TQM theory based on the study by [33], [27] suggested information quality management (IQM) framework which consists of 11 interdependent factors which are 1) IQM governance, 2) continuous IQM improvement, 3) training, 4) information quality requirements management, 5) information quality risk management, 6) information quality assessment/monitoring, 7) continuous information quality improvement, 8) information product lifecycle management, 9) storage management, 10) information security management, and 11) information architecture management. Furthermore, [28] enriched the work of [24], [25], [27] by suggesting the top three factors influencing data quality management, namely, 1) data governance, 2) management commitment and leadership, and 3) continuous data quality management improvement.

In the conclusion, the advancement of data quality study, ranging from specific data domain to general data domain provides a sound foundation in understanding and having deeper insight on issues related to data quality.

B. Master Data Quality

Acknowledged as an important asset and representing the core business process, assuring high-quality master data has gained extensive attention in the literature [39], [40], [45]. Concerning the improvement of master data quality, understanding factors influencing the quality of the data is a pre-requisite. Even though literature focusing on factors influencing master data quality is scarce, partial contribution by a few scholars such as [1], [21], [46], [47] providing a good starting point.

The first serious discussion and analyses of factors influencing master data quality were performed by [46], emphasizing that issues related to master data quality not only confined to technological aspects but more to organizational. Grounded on previous data quality theoretical foundation study by [9], [48]–[51], the author empirically validated five factors influencing master data quality which are the 1) delegation of responsibilities, 2) rewards, 3) data control, 4) employee competencies, and 5) information system.

As the continuation, a substantial work performed by [47] proposing 12 factors influencing master data quality which are 1) responsibilities for specific types of master data, 2) roles concerning data creation, use and maintenance, 3) organizational procedures, 4) management focus concerning data quality, 5) data quality measurements, 6) reward and reprimand about data quality, 7) training and education of data users, 8) written data quality policies and procedures, 9) emphasis on the importance of data quality by managers, 10) IT system for data management, 11) possibilities for input in existing IT system, and 12) usability of IT system. The identified factors were empirically validated using a survey mechanism that involved 787 Danish manufacturing company. The main difference in the work by [46] and [47] is the latter reclassified the factors identified in the previous literature to enable a more systematic understanding of the issues related to master data quality in ensuring the right improvement strategy.

A more specific perspective has been adopted by [1] that explored the challenges and requirements in managing master data quality in the context of digitalization. The author has adopted the SLR approach in getting a deeper insight into the current state of master data quality study and further validated the finding using 33 semi-structured interviews. In assuring the quality of master data during information sharing, the author suggested functional requirement for master data quality management (MDQM) tool that composed of six modules which are 1) analysis, 2) cockpit, 3) data model, 4) rules engine, 5) software architecture, and 6) software ergonomics. The functionality of each module can assist the organization in developing a tool for managing master data quality.

On another note, the study by [21] provides an understanding that different class of master data, exhibit distinct data quality challenges and requirements. The finding demonstrated the need to consider the development of a master data quality management approach based on the individual classes of master data. The author proposed a data quality assessment and improvement model that consists of eight elements which are 1) data quality assessment and improvement process, 2) technology, 3) protocol, 4) performance, 5) policy, 6) data standard, 7) data governance, and 8) data quality dimension.

Overall, although extensive research has been carried out in the field of master data quality supported by empirically validated finding, no single study exists that adopt both TQM as a theoretical lens and SLR as methodology. Theory helps in providing a systematic understanding of the real-world phenomenon, particularly provides a focus for the research [52]–[55]. In the case of data quality study, the wide adoption of TQM theory in understanding issues related to data quality is evidenced in many seminal works but, deficient in the context of master data quality. In the context of SLR methodology adoption, only evidenced in [1]. Nevertheless, the study by [21], [46], [47] does not systematically review all the relevant literature in discussing factors related to master data quality.

As a result of the lack of theoretical lens and systematic methodology, only partial contribution can be found in master data quality studies. In particular, finding by [1] emphasized on technological factors, while [21], [46], [47] unable to provide adequate and sufficient explanation on the master data quality challenges.

III. METHOD

SLR is a research method that provides a more structured and rigorous process in identifying and analyzing previous literature based on the specified research question. Normally, SLR-based study required the adaptation of established standards in guiding the researcher to perform the related and necessary process that will enable them to evaluate and examine the quality and rigor of a review. Therefore, this study is performed based on the guideline proposed by [56] that is designed particularly for Information System research, which consist of four main stages namely 1) planning, 2) selection, 3) extraction, and 4) analyses of findings. Each stage will be described further in the next section.

A. Planning

The planning stage emphasizes the identification of the research questions based on the study objective that acts as a frame in scoping the literature search. The main objective of this study is to investigate the factors influencing master data quality at the organizational level. Thus, this study formulated research questions which are 1) what are the factors influencing master data quality in the organization?, and 2) how do the factors influencing master data quality differ from other data domains?.

B. Selection

The selection stage identifies several relevant articles for the current study consist of three main processes. The first process is identifying the source of articles, followed by the construction of keywords, and lastly identification of inclusion criteria.

1) Source: The searching process covers seven main database sources, namely, 1) Web of Science, 2) Scopus, 3) ACM Digital Library, 4) Emerald, 5) Science Direct, 6) Springer Link, and 7) IEEE. Additionally, the study also includes Google Scholar to find more related articles on master data quality topics. The selection of databases was based on its coverage relating to information management source, expert recommendation, and accessibility of the database. The title, abstract, and keywords were used to conduct searches for journals, and proceedings, books, book chapters, and industry research.

a) Keywords: Construction of search keywords involves the process of 1) identification of alternative spellings and synonyms for major terms based on the thesaurus, dictionaries, encyclopedia, and past researches, 2) identification of keywords in relevant papers or books, and 3) usage of the Boolean OR to incorporate alternative spellings and synonyms [57]. Search keywords were constructed to retrieve as many articles as possible related to master data quality, the topic of interest in this study.

The search keywords are formulated by mentioning both the terms “master data quality” and “master information quality” due to the previous research in data management used both terms interchangeably. Search keyword also includes the term “master data management”, in reflection to the previous literature that referred master data management in relevance to the approach in managing master data. Thus, based on the search keywords, the initial search strings are (“master data quality”), (“master information quality”), and (“master data management”). Then, the search strings were joint using “OR” Boolean. The search strings were then used as the input to each electronic database to retrieve the articles based on the titles, abstracts, contents, and keywords, depending on the advanced search facility.

2) Inclusion criteria: The inclusion criteria are defined as means to reduce the number of studies to a certain amount that is reasonable to the author. There are three inclusion criteria formulated which are 1) language, 2) literature type, and 3) timeline as per Table I. In the first criteria, this study only focuses on the article that is written in the English language. The second criteria, limit the articles that are categorized only under journal, proceedings, books, and book chapters. Moreover, only articles between 2015 and 2020 are selected. Overall, a total of 2117 articles were found during the initial search, and 1285 articles were excluded based on exclusion criteria.

C. Extraction

A total of 832 articles were extracted for the third stage known as the study extraction. The manual searching process from Google Scholar is performed, in the case where the articles were not indexed in the selected database. The manual search resulted in additional two articles making the total articles 834. The metadata for the selected article include 1) title of the article, 2) publication year, 3) author, 4) abstract, 5) keywords, 6) article type, and 7) DOI/ISBN/ISSN Number is extracted. Then, the deduplication process is performed to remove the duplicated copies of the identified articles that exist across electronic repositories [58]. From this exercise, a total of 111 articles were removed during the checking of duplication, while 723 articles were further screened based on quality assessment criteria decided by the researcher.

At this stage, quality assessment was conducted by performing the practical screening against the 723 identical articles. Practical screening is the activity of screening the title and abstract of the articles based on quality assessment criteria to check the relevance of the articles [56]. The quality assessment criteria are 1) focus of the article, 2) mentioning any factor influencing master data quality, and 3) adequately describe the factors involved as per Table II. Consequently, a total of 708 articles were excluded because they are not fulfilling the quality assessment criteria. Finally, a total of 15 remaining articles are ready to be analyzed.

D. Analyses

This stage further analyzed 15 selected articles in answering the research questions. The detailed analyses are presented in the following Section IV.

TABLE I. INCLUSION AND EXCLUSION CRITERIA

Criteria	Inclusion	Exclusion
Language	English	Non-English
Article type	Research article, conference proceeding, book chapter, and book	Not categorized as a research article
Timeline	Between 2015 and 2020	Less than 2015

TABLE II. QUALITY ASSESSMENT CRITERIA

Code	Criteria
QA1	Is the main focus of the article is master data quality?
QA2	Are the articles describing any factor influencing master data quality?
QA3	Are the factors influencing master data quality adequately defined?

IV. RESULT

The systematic review process produced 15 related studies as presented in Table III. Regarding the credibility of the source, eight studies are from indexed journals [1], [14], [21], [59]–[63], four studies are from established conferences [13], [64]–[66], and three studies are from book publications [45], [67], [68]. In the case of present study, four articles were published in 2019 [14], [62], [63], [66], two articles in 2018[13], [65], four articles in 2017 [1], [59]–[61], two articles in 2016 [21], [64] [24, 86], and three articles in 2015 [45], [67], [68].

TABLE III. LIST OF RELATED ARTICLE BY YEAR

Year	Author	Source
2015	[67]	Apress
2015	[45]	epubli GmbH
2015	[68]	Morgan Kaufmann
2016	[21]	International Journal of Business Information Systems
2016	[64]	24th European Conference on Information Systems (ECIS 2016)
2017	[59]	Studies in Health Technology and Informatics
2017	[60]	Journal of Theoretical and Applied Information Technology
2017	[1]	Lecture Notes in Business Information Processing, Springer, Cham.
2017	[61]	Journal of Enterprise Information Management
2018	[13]	26th European Conference on Information Systems (ECIS 2018)
2018	[65]	International Conference on Information Management and Technology (ICIMTech)
2019	[14]	International Journal of Information Management
2019	[66]	International Conference on Smart Applications, Communications and Networking (SmartNets 2019)
2019	[62]	International Journal of Business Information Systems
2019	[63]	International Journal of Information Management

The detailed finding of the study is described based on the research questions.

A. RQI: What are the Factors Influencing Master Data Quality in the Organization?

Further analyses of the finding produced a total of 19 factors influencing master data quality, then the identified factors are further classified into five components which are organizational, managerial, stakeholder, technological and external as suggested by [25], [43], [44]. The theoretical perspective of the classification is useful to group the factors into specific components to have a broader overview of their effect on master data quality and allowing systematic analysis of the finding. As exhibited in Table IV, the five components are organizational (five factors), managerial (six factors), stakeholder (four factors), technological (two factors), and external (2 factors). Based on Table IV, the most frequently discussed factor is data governance which is mentioned in 11 out of 15 studies, followed by information system and data quality policy and standard which is discussed in more than half of the studies. It is then followed by data quality assessment, integration, continuous improvement, teamwork, data quality vision and strategy, understanding of the systems and data quality, data architecture management, and personnel competency with the occurrence between 4 and 7.

Lastly, with a frequency of less than 4, the factors are top management support, business driver, legislation, information security management, training, change management, customer focus, and data supplier management.

1) Organizational: Organizational is one of the components that have a major influence on master data quality. In particular, an organization does not only provide strategic direction to enable the implementation of a feasible road map in improving master data quality but also in many ways materialized the commitment in ensuring the achievement of data quality goals. In this case, a total of 11 studies were found focusing on an organizational component in improving master data quality. The discussed factors are data governance [14], [21], [68], [45], [59], [61]–[65], [67], teamwork [59], [61], [64], [67], data quality vision and strategy [45], [62], [63], [67], training [59], and change management [64].

TABLE IV. FACTORS INFLUENCING MASTER DATA QUALITY

Component	Factor	Author
Organizational	Data governance	[14], [21], [68], [45], [59], [61]–[65], [67]
Organizational	Teamwork	[59], [61], [64], [67]
Organizational	Data quality vision and strategy	[45], [62], [63], [67]
Organizational	Training	[59]
Organizational	Change management	[64]
Managerial	Data quality policy and standard	[14], [21], [59], [61], [62], [65], [67], [68]
Managerial	Data quality assessment	[21], [45], [59], [62], [64], [65], [68]
Managerial	Continuous improvement	[21], [45], [59], [62], [64]
Managerial	Understanding of the systems and data quality	[45], [59], [61], [64]
Managerial	Data architecture management	[45], [65]–[67]
Managerial	Information security management	[14], [68]
Stakeholder	Personnel competency	[14], [59], [64], [66]
Stakeholder	Top management support	[14], [64], [66]
Stakeholder	Customer focus	[62]
Stakeholder	Data supplier management	[1]
Technological	Information system	[1], [13], [14], [21], [45], [62], [63], [66]–[68]
Technological	Integration	[13], [62], [63], [66]–[68]
External	Business driver	[14], [67], [68]
External	Legislation	[14], [61], [67]

a) Data governance: Data governance involves the establishment of an organizational structure for managing master data quality that can be either a newly formed committee or reoccupied existing formal organizational structure. The latter is preferred to avoid any bureaucracy [45]. The core component of effective data governance is explained by the enactment of roles, responsibilities, and decision areas related to master data quality management [14], [21], [68], [45], [59], [61]–[65], [67]. Roles and responsibilities can be defined based on three organizational levels which are strategic, managerial, and operational [67]. The strategic level involves the role and responsibilities of the business sponsor, chief information officer (CIO), and chief operating officer (COO) which are the head of the IT and business department, and the leader for data governance. While managerial level includes the roles and responsibilities of the program manager and solution architect for the respective master data quality management initiative. Lastly, the operational level comprises the roles and responsibilities of the technical and business team.

Another strategy in defining roles and responsibilities in managing master data quality is through the concept of ownership [61]–[64]. Vilminko-Heikkinen and Pekkola [63] further explained the approach using three-level of master data quality management which are managerial level that involves the concept owner role responsible for the whole master data quality management initiative, support function role involving technical task and data domain level role consist of data owner task responsible for the data domain as a whole.

Apart from that, the roles and responsibilities must be determined not only for the internal decision area but must include the external process especially when the organization is involved with outsourcing activity [45]. Furthermore, in assuring the continuous quality of master data, roles and responsibilities at every stage of the data life cycle such as data creation, modification, access, and deletion should be defined [45], [59]. [45], [59].

b) Teamwork: Sufficient communication, understanding, and involvement between technical and business employees across the department are the provision of effective teamwork [24]–[26], [28]. The effective teamwork can be facilitated using business friendly approach [69]. The management of master data quality must involve both technical and business people to ensure fair and equal accountability [59], [61], [64], [67]. Furthermore, it is essential to strengthen the alignment of responsibilities between both parties [59] especially at the high-level coordination [61] involving the enforcement of policies to support business activities and also compliance to regulation [67]. Lack of teamwork effort normally leaves the task of managing master data to the technical people [64] or worst still to no man island and could potentially compromise the quality of master data.

c) Data quality vision and strategy: Data quality vision and strategy provide the direction in ensuring the quality of master data [45], [62], [63], [67]. Data quality vision is developed in line with the organization's vision, providing the key business initiatives to support the organization's vision [67]. While data quality strategy is the detailed component of the data quality vision, elaborating the business case, and roadmap for the implementation of data quality initiative [63], [67].

d) Training: Effective and adequate training is essential in ensuring the employees are equipped with sufficient knowledge and skills in managing master data quality [24], [25], [27], [28], [70]. According to [59], in the case of the industry that deals with specific technical data such as the health field, employees need to have sufficient training not only on how to properly perform data entry and data processing, but capable to perform data quality checking especially involving semantic data quality checking to ensure the quality of the master data.

e) Change management: Change management refers to the organization's capability in managing internal and external change such as merging, technology transformation, government regulations, and market shift [24], [25], [28]. According to [64], change can be managed using a top-down approach or bottom-up approach, with the ultimate goal is to ensure the commitment and involvement of the employee in taking up new responsibilities.

2) Managerial: Improving master data quality requires effective and efficient management involving the provision of comprehensive data quality guidelines in ensuring the process of managing master data quality properly performed. As previously mentioned, a total of 11 studies were found to focus on managerial component related to master data quality improvement. Nevertheless, the analyses for this component has resulted in a total of six factor, namely data quality policy and standard [14], [21], [59], [61], [62], [65], [67], [68], data quality assessment [21], [45], [59], [62], [64], [65], [68], continuous improvement [21], [45], [59], [62], [64], understanding of the systems and data quality [45], [59], [61], [64], data architecture management [45], [65]–[67] and information security management [14], [68].

a) Data quality policy and standard: Policy and procedure act as a frame to enable the improvement of data quality that includes the data that is managed internally by the organization [71], [72]. On top of that, data quality policy and standards provide managerial level guidance in implementing master data quality management initiatives. According to [24], [25], [28], the guideline normally has two main parts namely, what to achieve concerning data quality goal and how to achieve the stipulated goal [24], [25], [28].

In particular, data quality policy and standards spell out the detailed definition of master data [59], [62] and master data quality management taxonomy [61]. The definition includes the structure of the data, a business process that uses specific master data, the reason the master data is created, and governance of the master data [59], [61], [62], [67]. Other than that, the document also contains business rules for managing master data quality, guidelines in responding to data quality issues, and service level agreement (SLA) that act as data quality indicators [21], [59]. Ultimately, a well-written data quality policy and standard must support the business process

of the organization including compliance to regulation, fulfilling customer needs, and providing consistent reporting [14], [67], [68].

b) Data quality assessment: Data quality assessment is a pre-requisite step before proceeding with any data quality improvement initiative [27]. Several seminal authors emphasized that "only what can be measured can be improved" [39], [40], [49]. Therefore, measurement of data quality has to be done to determine the level of data quality over time. Data quality assessment consists of four main phases namely, definition, measurement, analysis, and improvement, which involve various stakeholders such as data collector, data custodian, and data user [34], [73].

During the definition phase, analysis of the current state of data quality is performed to discover any problem related to data quality [21], [65], then data quality requirement and measurement metric is determined based on a key performance indicator (KPI) for data quality and also business process performance [45], [64], [68]. Later, the data quality dimension is identified [21], [62], which acts as a facet for data quality that will be used in the subsequent phase. In the case of identifying data quality dimensions, the seminal work by [73], [74] are most frequently cited.

Then, measurement of data quality is performed which involves quantitative and qualitative strategy [28] based on identified KPI. The measurement can include syntax and semantic checking utilizing current technology such as rules engine and fuzzy search [45], [59]. As for the analysis phase, the assessment result should be compared to the earlier defined data quality requirement which is based on KPI, perform benchmarking, and prioritize data quality improvement strategy [21], [45], [59]. Finally, the improvement of data quality is implemented with adequate monitoring in place [21].

c) Continuous improvement: Continuous improvement of master data quality is normally driven by the changes in the internal and external environment such as technology and regulation that requires a proper response by the organization [24], [25]. Continuous improvement is an ongoing process for assuring the quality of master data, which involves the implementation of a preventive measure focusing more on the business process betterment [26], [59]. In particular, involving the installation of data quality elements in every phase of the data life cycle is required that include data collection, processing, deletion, and archiving [21], [45].

Another aspect of continuous improvement involves the assessment of the maturity level of master data quality management practice in the organization [64]. On top of that, regular data quality examination is also required, not only to gauge the level of master data quality but to enable the employee to see the effect of their work on data quality [59]. Finally, the result of the data quality improvement initiative must be integrated into the organization's operations for reporting and monitoring purposes [45], [59].

d) Understanding of the systems and data quality: Improvement of master data quality requires the understanding of how the application works, the importance of data quality and the relation with the business objective, and also the usefulness of the data [24], [25], [28]. Employees should understand the effect of poor data quality [45], [59] on the organization and be aware that the management of master data quality is an enterprise-wide initiative, which does not only affect specific business units but the organization as a whole [61], [64]. The understanding is important in motivating the employees to give full commitment in improving master data quality.

e) Data architecture management: Data architecture management involves the coordination of business process, application, data, and integration process [27] that includes the definition of global and local data, retention, and distribution of data [66], [67]. Master data quality requirement provides a basis for the data architecture management [65] include the identification of required tools and technology to build the solution [67].

f) Information security management: Information security management is referred to as the extent of the process and practice in the organization to safeguard the confidentiality of the master data [27], [28]. According to [14], [68], the privacy and security of master data include the protection from unauthorized access and the provision of reliable and secure communication means during data sharing.

3) Stakeholder: In this section, it is important to gain a deeper understanding of the influence of stakeholders as one of the critical components in managing master data quality. As previously mentioned, a total of six studies focused on master data quality related to stakeholders. The current study, managed to further categorized the component into four factors namely personnel competency [14], [59], [64], [66], top management support [14], [64], [66], customer focus [62], and data supplier management [1].

a) Personnel competency: Managing master data quality requires the employees to be equipped with sufficient skills and knowledge in both technical and business areas. According to [64], the organization should have a clear definition of the knowledge and skills that are needed for managing master data quality to ensure the right people are employed for the right task. Furthermore, according to [14], [59], [64], [66], the organization should appoint well-trained, experienced, and qualified personnel in both technical and business areas representing all departments in the organization.

b) Top management support: Awareness, competency, and leadership on master data quality possessed by top management is another important factor that is frequently discussed in the previous studies. Top management must be aware of the importance of master data quality and support activity related to master data quality management [64]. Apart from that, managing master data quality requires well-trained personnel, hence, top management should provide sufficient resources in improving skills and knowledge [14]. According to [66], top management also should focus on rewards and recognition programs for employees within an organization.

c) Customer focus: Focus on the user's needs is important to ensure the quality of data satisfies the defined requirement [24], [25], [27], [28]. Users must be involved during the data quality requirement elicitation, to ensure the correct requirement is captured [62]. In the context of system development process, data model that uses business metadata such as Source-Transaction-Agent (STA) model can be utilized to assist business and IT person to communicate and participate effectively and efficiently in business data modelling [69].

d) Data supplier management: Data supplier refers to the party, either internal and external to the organization that provides raw, unorganized data, while data supplier management is defined as having an effective relationship with the data provider by having an agreement about the acceptable level of data quality supplied and provide regular data quality reporting and technical assistance to data suppliers [24], [25], [28]. According to [1], sata provider is responsible to provide quality master data with fewer errors.

4) Technological: The technological component is considered as the operational level of the master data quality management initiative [75]. As previously mentioned, a total of 10 studies focused on master data quality related to technology. The present study managed to further classified the component into two factors which are information system [1], [13], [14], [21], [45], [62], [63], [66]–[68] and integration [13], [62], [63], [66]–[68].

a) Information system: In supporting the effective management of master data quality, the information system should provide sufficient functions, cutting edge architecture, and also adequate ergonomics features. The system should have the capability in assuring the quality of master data such as data profiling, data cleansing, data matching, data merging, data synchronization, and data consolidation [62], [63], [67], [68]. On top of that, [1] highlighted that the system should be capable to perform testing and simulation of data quality measurement and provide data quality monitoring in the form of a cockpit.

Furthermore, a system architecture is developed based on business process architecture [62] and adheres to the modular principle, adaptability, and reconfigurability [1]. Lastly, [1] explained that the system also must possess adequate ergonomic features such as easy to use, understandable, and comprehensive.

b) Integration: Since master data can exist in multiple sources, a certain degree of data integration is needed to preserve the quality of master data. Data integration implementation depends on the requirement such as data volume, data latency, nature of data, and the number of staging layers needed [67]. According to [68], the Entity Identity Information Management (EIIM) approach can be used to maintain the integrity of the master data, which is the fundamental element for Master Data Management (MDM).

In addition to that [13] proposed the usage of a federated approach to integrate the data based on shared attributes metadata to overcome the problem caused by a single source of truth approach.

5) External: External component refers to the factor that affects the master data quality which is not within the organization's control but somehow needs to be faced by the organization to stay competitive or comply with the regulation. As previously mentioned, a total of four studies focused on the master data quality related to the external component. The present study managed to further classify the component into two factors which are business driver [14], [67], [68] and legislation [14], [61], [67].

a) Business driver: In order to stay competitive and relevant in the data-driven economy, requires the organization to effectively and efficiently adapt to the ever-changing business need that in many ways require changes in how master data is managed. Poorly managed data, affect the data quality, hence influence the organization's performance. The related business driver that should be considered by organizations includes consumer demand for higher quality product or service, capability in offering new product and services in less time, single view reporting to enable more informed decision making, data integration from multiple sources, return on investment and ensuring data security especially during data sharing process [14], [67], [68].

b) Legislation: Every organization operates in an environment that is governed by certain rules and regulations that have to adhere. Such legislation includes the data protection act in ensuring the confidentiality and privacy of customer data is assured [14], [61], [67]. Failure to comply with the stipulated legislation not only affects the reputation of the organization, but to make it worse, is the possibility to face a lawsuit.

B. RQ2: How do the Factors Influencing Master Data Quality Differ from other Data Domain?

Table V summarized the differences for the factor that influence master data quality compared to other domains of data namely accounting data, health data, and general data.

Based on Table V, there are a total of 34 factors that influence accounting data, health data, general data, and master data with some differences. In particular, the factor that discussed only in master data is business driver, while in contrast 15 factors are discussed in other data domain but not explicitly in the master data domain namely 1) organizational structure, 2) organizational culture, 3) performance evaluation and rewards, 4) evaluate cost/benefit tradeoffs, 5) physical environment, 6) risk management, 7) storage management, 8) usage of data, 9) internal control, 10) input control, 11) staff participation, 12) middle management's commitment, 13) role of data quality and data quality manager, 14) audit, and 15) personnel relation. On top of that, it is worth highlighting that only 6 factors are discussed across all data domains, which include 1) teamwork, 2) data quality policy and standard, 3) continuous improvement, 4) top management support, 5) information system, and 6) integration.

TABLE V. FACTORS INFLUENCING THE QUALITY OF ACCOUNTING DATA, HEALTH DATA, GENERAL DATA, AND MASTER DATA

Factor	A	B	C	D
Business driver				/
Organizational structure	/			
Organizational culture	/		/	
Performance evaluation and rewards	/		/	
Evaluate cost/benefit tradeoffs	/		/	
Physical environment	/		/	
Risk management	/		/	
Storage management			/	
Usage of data	/			
Internal control	/			
Input control	/		/	
Staff participation		/		
Middle management's commitment	/			
Role of data quality and data quality manager	/			
Audit	/		/	
Personnel relation	/		/	
Teamwork	/	/	/	/
Data quality policy and standard	/	/	/	/
Continuous improvement	/	/	/	/
Top management support	/	/	/	/
Information system	/	/	/	/
Integration	/	/	/	/
Training	/		/	/
Change management	/		/	/
Data quality assessment	/		/	/
Understanding of the systems and data quality	/		/	/
Personnel competency	/		/	/
Customer focus	/		/	/
Data supplier management	/		/	/
Data governance			/	/
Data architecture management			/	/
Information security management			/	/
Data quality vision and strategy	/			/
Legislation	/			/

(A) Accounting data B) Health data C) General data D) Master data

While other factors namely data governance, data architecture management, information security management, data quality vision and strategy training, change management, data quality assessment, understanding of the systems and data quality, personnel competency, customer focus, personnel relation, data supplier management, and legislation are mentioned in master data quality study but discussed partially in another study. Overall, the similarity and dissimilarity of the finding provide a good justification in pursuing further study on master data quality.

V. DISCUSSION

The result of the analyses in Table IV gives further insight into the potential factor that could impact master data quality. It should also be noted that the first three factors which are data governance, data quality policy and standard, and information system clearly stand out from the rest. The high occurrence of data governance is noteworthy, 11 studies in master data discuss the factor, while only two studies in other data domains highlight the factor [27], [28]. This outcome is probably due to the increasing importance of data to organizations, particularly in the context of digital transformation, which has given rise to the need of establishing the roles and responsibilities in managing master data quality such as data ownership [61], [63], [64], among other.

However, establishing roles and responsibilities for managing master data quality exhibits complex challenges, since master data do not belong to a specific department but is an asset for the organization as a whole. Hence, managing master data requires an enterprise-wide approach compared to other data domains that are more compartmentalized to a specific business unit. The responsibilities are huge, where employees are reluctant to carry such accountability. The organization also might find difficulties in shifting the data management approach from department-based to enterprise wide-based. Hence, the complexity creates the need to further study in getting more insight and understanding of the phenomena.

Based on Table V, it can be summarized that there are differences in the factors influencing master data quality compared to other data domains. As highlighted by [21], different data domains, possess different challenges and requirements, hence requiring a more tailored suit management approach in ensuring data quality. In particular, the business driver factor is discussed in three master data studies domains but none in other data domains. The reason is might due to the effect of digital transformation to the organization that requires more proactive action in managing master data quality, while other data domain does not consider it as a threat.

Another interesting result to explore is regarding the total of 15 factors that are not explicitly discussed in the master data domain but mentioned in another data domain namely organizational structure, organizational culture, performance evaluation and rewards, evaluate cost/benefit tradeoffs, physical environment, risk management, storage management, usage of data, internal control, input control, staff participation, middle management's commitment, the role of data quality and data quality manager, audit, and personnel relation. As for the organizational structure, the factor was consolidated under the data governance factor due to the relevancy and suggestion by [28], whereas expanding the definition of data governance. Hence, the rest of the factor needs further investigation regarding the relevancy in the master data context.

Overall, the analyses suggested a total of 19 factors are relevant in the context of master data quality. The cross-reference of the identified 19 factors against the study in other

domain shows that teamwork, data quality policy and standard, continuous improvement, top management support, information system, integration, training, change management, data quality assessment, understanding of the systems and data quality, personnel competency, customer focus, and data supplier management are considered as established factors as were discussed in a minimum of three data domains of the study including master data domain. Meanwhile, as for the data governance factor, important to note that, even though lack of discussion of the factor in other domains of study, but the emphasis is so obvious in the master data domain, making it the most dominant factor. Other than that, data architecture management, information security management, data quality vision and strategy, and legislation discussed in two domains including the master data domain, which surely require a deeper understanding. The overall result, reflect the need and further justifying the need to pursue a study in master data quality to get a deeper insight into the influencing factors.

VI. Conclusion

A better understanding of the factors influencing master data quality will enable practitioners to improve master data quality. There is evidence that the effect of digital transformation requires the organization to change how it manages master data. Poorly managed master data, produce low-quality data and affect organization performance in term of fulfilling increasing customer demand, providing 360-degree single view reporting and integration of multiple data sources, to name a few.

Even though master data is acknowledged as an asset to the organization and a core element to the business process, the comprehensive study on factors influencing master data quality is very limited. With the aim to diminish the gap, this study can be considered as one of the first attempts to thoroughly review factors influencing master data quality.

The significant findings that transpired from this review study are that 19 factors of master data quality have been identified and categorized into five components which are organizational, managerial, stakeholder, technological, and external. The top 10 most influential factors are data governance followed by information system, data quality policy and standard, data quality assessment, integration, continuous improvement, teamwork, data quality vision and strategy, understanding of the systems and data quality, and data architecture management.

Interestingly, the analyses show that there are some dissimilarities for factors influencing master data quality, compared to other data quality domains. In the domain of accounting, the differences include organizational structure, organizational culture, performance evaluation and rewards, evaluate cost/benefit tradeoffs, physical environment, risk management, usage of data, internal control, input control, middle management commitment's, role of data quality and data quality manager, audit, personnel relation, business driver, data governance, data architecture management, and information security management.

While for the health data, the contradict factors involve staff participant, business driver, training, change management, data quality assessment, understanding of the systems and data quality, personnel competency, customer focus, data supplier management, data governance, data architecture management, information security management, data quality vision and strategy, and legislation. Eventually, for general data, the differences include organizational culture, performance evaluation and rewards, evaluate cost/benefit tradeoffs, physical environment, risk management, storage management, input control, audit, personnel relation, business driver, data quality vision and strategy, and legislation. It is recommended to further investigate these factors using an in-depth interview to better understand the phenomenon.

Acknowledgment

The study was financially supported by the Research Grant ETP-2013-060, Universiti Kebangsaan Malaysia.

References

[1] T. Schäffer and C. Leyh, "Master data quality in the era of digitization - Toward inter-organizational master data quality in value networks: A problem identification," Piazolo F., Geist V., Brehm L., Schmidt R. Innov. Enterp. Inf. Syst. Manag. Eng. ERP Futur. 2016. Lect. Notes Bus. Inf. Process. Springer, Cham, vol. 285, pp. 99–113, 2017.

[2] Z. Mohammad Yusof, Pengurusan rekod dan maklumat: Isu dan cabaran. Penerbit Universiti Kebangsaan Malaysia, Bangi, Malaysia, 2015.

[3] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and framework for data and information quality research," ACM J. Data Inf. Qual., vol. 1, no. 1, pp. 1–17, 2009.

[4] M. Mukred and Z. M. Yusof, "The Delone–McLean information system success model for electronic records management system adoption in higher professional education institutions of Yemen," Lect. Notes Data Eng. Commun. Technol., vol. 5, pp. 812–823, 2018.

[5] S. Nelke, M. Oberhofer, Y. Saillet, and J. Seifert, "U.S. Patent 2015/0066987 A1: Method and system for accessing a set of data tables in a source database," 2015.

[6] M. Y. Choi, C. J. Moon, K. S. Park, and D. K. Baik, "An enterprise master data model based on the data taxonomy based on their origin," in International Conference on Enterprise Information Systems and Web Technologies (EISWT), 2010.

[7] A. Dreibelbis, E. Hechler, I. Milman, M. Oberhofer, P. van Run, and D. Wolfson, Enterprise master data management: An SOA approach to managing core information. Pearson Education, 2008.

[8] D. Loshin, Master data management. Morgan Kaufmann OMG Press, 2009.

[9] H. A. Smith and J. D. Mckeen, "Developments in practice XXX: Master data management: Salvation or snake oil?," Commun. Assoc. Inf. Syst., vol. 23, no. 4, pp. 1–11, 2008.

[10] F. Haneem, N. Kama, and A. Azmi, "Master data identification in public sector organisations," Adv. Sci. Lett., vol. 22, no. 10, pp. 2999–3003, 2016.

[11] J. S. Arlbjørn, C. Y. Wong, and S. Seerup, "Achieving competitiveness through supply chain integration," Int. J. Integr. Supply Manag., vol. 3, no. 1, pp. 4–24, 2007.

[12] B. Otto, K. M. Hüner, and H. Österle, "Toward a functional reference model for master data quality management," Inf. Syst. E-bus. Manag., vol. 10, no. 3, pp. 395–425, 2012.

[13] T. Dahlberg, A. Lagstedt, and T. Nokkala, "How to address master data complexity in in-formation systems development – A federative approach," in 26th European Conference on Information Systems (ECIS 2018), 2018, pp. 1–15.

[14] F. Haneem, N. Kama, N. Taskin, D. Pauleen, and N. A. Abu Bakar, "Determinants of master data management adoption by local government organizations: An empirical study," Int. J. Inf. Manage., vol. 45, pp. 25–43, 2019.

[15] T. C. Redman, "The impact of poor data quality on the typical enterprise," Commun. ACM, vol. 41, no. 2, pp. 79–82, 1998.

[16] A. Haug, F. Zachariassen, and D. van Liempd, "The costs of poor data quality," J. Ind. Eng. Manag., vol. 4, no. 2, pp. 168–193, 2011.

[17] W. W. Eckerson, "Data quality and the bottom line: Achieving business success through a commitment to high quality data," 2002.

[18] Royal Mail and DataIQ, "How better customer data drives marketing performance and business growth," 2016.

[19] C. Lehmann, K. Roy, and B. Winter, "The state of enterprise data quality: 2016: Perception, reality and the future of DQM," 2016.

[20] Deloitte, "Complexity: Overcoming obstacles and seizing opportunities: The Deloitte global CPO survey 2019," 2019.

[21] R. Silvola, J. Harkonen, O. Vilppola, H. Kropsu-Vehkapera, and H. Haapasalo, "Data quality assessment and improvement," Int. J. Bus. Inf. Syst., vol. 22, no. 1, pp. 62–81, 2016.

[22] A. Hayler, "Ten years on, master data management software market comes of age," 2014. .

[23] M. Treder, "Chapter 5- Masterdata management," in The chief data officer management handbook: Set up and run an organization's data supply chain, Apress, 2020, pp. 61–78.

[24] H. Xu, "What are the most important factors for accounting information quality and their impact on AIS data quality outcomes?," J. Data Inf. Qual., vol. 5, no. 4, pp. 1–22, 2015.

[25] H. Xu, A. Koronios, and N. Brown, "Managing data quality in accounting information systems," in IT-based management: Challenges and solutions, Idea Group Publishing: Hershey PA, 2002, pp. 277–299.

[26] C. Liu, D. Zowghi, and A. Talaei-Khoei, "An empirical study of the antecedents of data completeness in electronic medical records," Int. J. Inf. Manage., vol. 50, pp. 155–170, 2020.

[27] S. Baškarada and A. Koronios, "A critical success factor framework for information quality management," Inf. Syst. Manag., vol. 31, no. 4, pp. 276–295, 2014.

[28] A. Lucas, "Critical success factors for corporate data quality management," in World Conference on Information Systems and Technologies, 2019, pp. 630–644.

[29] E. Hassan, Z. M. Yusof, and K. Ahmad, "Factors affecting information quality in the Malaysian public sector," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 9, no. 1, pp. 32–38, 2019.

[30] E. Hassan, Z. M. Yusof, and K. Ahmad, "Modeling of information quality management in malaysian public sector: A pls-sem approach," J. Theor. Appl. Inf. Technol., vol. 96, no. 19, pp. 6361–6375, 2018.

[31] J. M. Juran and A. B. Godfrey, Juran's quality handbook, Fifth edit. McGraw-Hill, 1998.

[32] W. E. Deming, Out of the crisis. MIT Press, 1986.

[33] J. Motwani, "Critical factors and performance measures of TQM," TQM Mag., vol. 13, no. 4, pp. 292–300, 2001.

[34] R. Y. Wang, "A product perspective on total data quality management," Commun. ACM, vol. 41, no. 2, pp. 58–65, 1998.

[35] L. Bertossi and M. Milani, "Ontological multidimensional data models and contextual data quality," J. Data Inf. Qual., vol. 9, no. 3, pp. 1–36, 2018.

[36] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi, "Modeling information manufacturing systems to determine information product quality," Manage. Sci., vol. 44, no. 4, pp. 462–484, 1998.

[37] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," Manage. Sci., vol. 31, no. 2, pp. 150–162, 1985.

[38] T. C. Redman, Data quality for the information age. Artech House, 1996.

[39] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," Commun. ACM, vol. 39, no. 11, pp. 86–95, 1996.

[40] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," J. Manag. Inf. Syst., vol. 12, no. 4, pp. 5–34, 1996.

[41] G. K. Tayi and D. P. Ballou, "Examining data quality," Commun. ACM, vol. 41, no. 2, pp. 54–57, 1998.

[42] S. Watts, G. Shankaranarayanan, and A. Even, "Data quality assessment in context: A cognitive perspective," Decis. Support Syst., vol. 48, no. 1, pp. 202–211, 2009.

[43] H. Xu, "Data quality issues for accounting information systems' implementation: Systems, stakeholders, and organizational factors," J. Technol. Res., pp. 1–11, 2009.

[44] H. Xu, J. H. Nord, G. D. Nord, and B. Lin, "Key issues of accounting information quality management: Australian case studies," Ind. Manag. Data Syst., vol. 103, no. 7, pp. 461–470, 2003.

[45] B. Otto and H. Österle, Corporate data quality: Prerequisite for successful business models. epubli GmbH, 2015.

[46] A. Haug and J. S. Arlbjørn, "Barriers to master data quality," J. Enterp. Inf. Manag., vol. 24, no. 3, pp. 288–303, 2011.

[47] A. Haug, J. S. Arlbjørn, F. Zachariassen, and J. Schlichter, "Master data quality barriers: An empirical investigation," Ind. Manag. Data Syst., vol. 113, no. 2, pp. 234–249, 2013.

[48] A. Umar, G. Karabatis, L. Ness, B. Horowitz, and A. Elmagardmid, "Enterprise data quality: A pragmatic approach," Inf. Syst. Front., vol. 1, no. 3, pp. 279–301, 1999.

[49] L. P. English, Improving data warehouse and business information quality: Methods for reducing costs and increasing profits. Wiley & Sons, 1999.

[50] H. Xu, J. H. Nord, N. Brown, and G. D. Nord, "Data quality issues in implementing an ERP," Ind. Manag. Data Syst., vol. 102, no. 1, pp. 47–58, 2002.

[51] Y. W. Lee, L. L. Pipino, J. D. Funk, and R. Y. Wang, Journey to data quality. The MIT Press, 2006.

[52] C. Kivunja, "Distinguishing between theory, theoretical framework, and conceptual framework: A systematic review of lessons from the field," Int. J. High. Educ., vol. 7, no. 6, pp. 44–53, 2018.

[53] S. Reeves, M. Albert, A. Kuper, and B. D. Hodges, "Qualitative research: Why use theories in qualitative research?," BMJ, vol. 337, no. 7670, pp. 631–634, 2008.

[54] A. Tashakkori and C. Teddlie, Handbook of mixed methods in social and behavioral research. Sage Publications, 2010.

[55] L. Varpio, E. Paradis, S. Uijtdehaage, and M. Young, "The distinctions between theory, theoretical framework, and conceptual framework," Acad. Med., pp. 989–994, 2020.

[56] C. Okoli, "A guide to conducting a standalone systematic literature review," Commun. Assoc. Inf. Syst., vol. 37, no. 1, pp. 879–910, 2015.

[57] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[58] Q. He, Z. Li, and X. Zhang, "Data deduplication techniques," in International Conference on Future Information Technology and Management Engineering, FITME, 2010, vol. 1, pp. 430–433.

[59] K. Arthofer and D. Girardi, "Data quality- and master data management - A hospital case," Stud. Health Technol. Inform., vol. 236, pp. 259–266, 2017.

[60] F. Haneem, A. Azmi, and N. Kama, "Co-dependence relationship between master data management and data quality: A review," J. Theor. Appl. Inf. Technol., vol. 95, no. 22, pp. 6323–6335, 2017.

[61] R. Vilminko-Heikkinen and S. Pekkola, "Master data management and its organizational implementation: An ethnographical study within the public sector," J. Enterp. Inf. Manag., vol. 30, no. 3, pp. 454–475, 2017.

[62] R. Silvola, A. Tolonen, J. Harkonen, H. Haapasalo, and T. Mannisto, "Defining one product data for a product," Int. J. Bus. Inf. Syst., vol. 30, no. 4, p. 489, 2019.

[63] R. Vilminko-Heikkinen and S. Pekkola, "Changes in roles, responsibilities and ownership in organizing master data management," Int. J. Inf. Manage., vol. 47, pp. 76–87, 2019.

[64] R. Vilminko-Heikkinen, P. Brous, and S. Pekkola, "Paradoxes, conflicts and tensions in establishing master data management function," in 24th European Conference on Information Systems (ECIS), 2016, pp. 1–16.

[65] Z. Murti, A. Andarrachmi, A. N. Hidayanto, and S. B. Yudhoatmojo, "Master data management planning: (Case study of personnel information system at XYZ Institute)," in International Conference on

Information Management and Technology (ICIMTech), 2018, pp. 160–165.

[66] T. E. Hutang and B. M. Kalema, "The effects of demographic variables on master data quality management to improve service delivery," in International Conference on Smart Applications, Communications and Networking (SmartNets), 2019, pp. 1–6.

[67] S. Chaki, "Pillar No. 4: Master information management," in Enterprise information management in practice: Managing data and leveraging profits in today's complex business environment, Apress, 2015, pp. 63–78.

[68] J. R. Talburt and Y. Zhou, "The value proposition for MDM and big data," in Entity information life cycle for big data: Master data management and information integration, Morgan Kaufmann, 2015, pp. 1–16.

[69] I. Mohamed and M. F. Noordin, "STA data model for effective business process modelling," Procedia Technol., vol. 11, pp. 1218–1222, 2013.

[70] H. Xu, "Managing accounting information quality: An Australian study," in International Conference on Information Systems (ICIS), 2000, pp. 628–634.

[71] M. Basri, Z. Mohammad Yusof, and N. A. Mat Zin, Dasar maklumat nasional di Malaysia. Penerbit Universiti Kebangsaan Malaysia, Bangi, 2013.

[72] N. Abdul Halim, Z. M. Yusof, and N. A. M. Zin, "The requirement for information governance policy framework in Malaysian public sector," Int. J. Eng. Technol., vol. 7, no. 4.15, pp. 235–239, 2018.

[73] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: A methodology for information quality assessment," Inf. Manag., vol. 40, no. 2, pp. 133–146, 2002.

[74] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: Product and service performance," Commun. ACM, vol. 45, no. 4, pp. 184–192, 2002.

[75] J. Xiao, K. Xie, and X. Wan, "Factors influencing enterprise to improve data quality in information systems application - An empirical research on 185 enterprises through field study," in 16th International Conference on Management Science and Engineering, IEEE, Moscow, Russia, 2009, pp. 23–33.

Hybrid Feature Selection and Ensemble Learning Methods for Gene Selection and Cancer Classification

Sultan Noman Qasem[1], Faisal Saeed[2]
Computer Science Department, College of Computer and Information Sciences[1]
Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia[1]
Computer Science Department, Faculty of Applied Science, Taiz University, Taiz, Yemen[1]
Information Systems Department, College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia[2]

***Abstract*—A promising research field in bioinformatics and data mining is the classification of cancer based on gene expression results. Efficient sample classification is not supported by all genes. Thus, to identify the appropriate genes that help efficiently distinguish samples, a robust feature selection method is needed. Redundancy in the data on gene expression contributes to low classification performance. This paper presents the combination for gene selection and classification methods using ranking and wrapper methods. In ranking methods, information gain was used to reduce the size of dimensionality to 1% and 5%. Then, in wrapper methods K-nearest neighbors and Naïve Bayes were used with Best First, Greedy Stepwise, and Rank Search. Several combinations were investigated because it is known that no single model can give the best results using different datasets for all circumstances. Therefore, combining multiple feature selection methods and applying different classification models could provide a better decision on the final predicted cancer types. Compared with the existing classifiers, the proposed assembly gene selection methods obtained comparable performance.**

Keywords*—Microarray; gene selection; ensemble classification; cancer classification; gene expression

I. Introduction

Gene expression is called the process of transcription of the Deoxyribo Nucleic Acid (DNA) sequence into Ribo Nucleic Acid (RNA). The expression frequency of a gene shows the average number of copies of the cell-produced RNA in that gene and is associated with the corresponding volume of protein [1].

Microarray is the technique for simultaneous measurements of the expression level in a single chip of tens of thousands of genes. Microarrays therefore provide an effective way to collect data that can be used to establish the pattern of expression of thousands of genes. In most classification issues, high gene expression data is a major challenge. Therefore, not all genes also lead to cancer. A broad variety of genes have no clinical importance or insignificance. However, incorrect diagnosis can also be accomplished by using both genes in the Microarray classification of gene expression. The two key explanations for low classification precision are two: large number of features (genes) against limited sample size and dimensional consistency in articulated data [2]. Subsequently, the decrease in dimensions is necessary. Standard machine learning methods have not been effective, since these methods are better suited when there are more samples than features.

In order to solve these problems, selection algorithms for dimension reduction or features (gene) were used. The gene selection methods are usually divided into three groups, namely filter, wrapper and embedded methods. The filter procedure requires the individual evaluation of each feature using its statistical characteristics in general. The wrapper approach uses training strategies to choose the best subset of features. By the precision of the particular classifier the efficiency of the wrapper technique is calculated. In the wrapper method evolutionary or bio-inspired algorithms are also used to direct the search process. The embedded approach aims for the best feature subset and is implemented in the classification scheme. The general structure for feature selection was recently complemented with hybrid and ensemble approaches. The filter and the wrapper approaches are designed to take advantage of hybrid. Extensive works have investigated this issue and proposed several methods such as [3-16].

Several feature selection methods have been applied. For instance, the authors in [17-19] proposed hybrid methods to combine filter and wrapper algorithms to overcome the disadvantage of each individual one. Conventional optimization algorithms are not efficiently working in the feature selection of large scale problems [20]. Alternatively, different meta-heuristic algorithms have been adapted for feature selection issues. Examples of these algorithms are Genetic Algorithm (GA) [21], Ant Colony Optimization [22], Simulated Annealing [23], and Particle Swarm Optimization (PSO) [24, 25]. In addition, a modified support vector machine (SVM) was also suggested to select the minimum possible genes [26]. Multi-objective version of bat algorithm for binary feature selection [27] and Genetic Bee Colony (GBC) algorithm [28] were successfully utilized in high dimensional datasets. Moreover, a hybrid feature selection algorithm was proposed that combines the mutual information maximization (MIM) and the adaptive genetic algorithm (AGA) [19]. The reduced gene expression dataset presented higher classification accuracy compared with conventional feature selection algorithms.

In addition, a binary version of Black Hole Algorithm called BBHA was proposed for solving feature selection problem in biological data. However, the tested classifiers were under tree family, and other kinds of classifiers were not assessed [29]. Along this line, the assessment of different classifiers such as artificial neural network (ANN) [30] and

fuzzy decision tree algorithm [31] has been made upon microarray data. In addition, the two evolutionary algorithms of PSO and GA are usually used in wrapper form [17, 20]. PSO is known to be a memory enabled algorithm compared with other algorithms, it requires few parameters to be adjusted, so it is simple and efficient [18, 32]. Kar et al. [33] proposed a PSO–adaptive K-nearest neighbors (KNN) based gene selection method and they used a heuristic for selecting the optimal values of K, while the classification accuracies have been tested using SVM algorithm. Furthermore, Jain et al. reported a two phase hybrid model for cancer classification, integrating Correlation-based Feature Selection (CFS) with improved-Binary Particle Swarm Optimization (iBPSO) using Naive-Bayes as the only classifier [34].

Moreover, Almutiri and Saeed [35], proposed a new combination for gene selection that utilized Chi Square and SVM Recursive Feature Elimination. This proposed method was called ChiSVMRFE and considered as ranking method. The top 10% of the genes were selected based on the high obtained weights and then SVM-RFE was used to remove the genes with lower weights. Only 10 features were selected and fed to several machine learning methods such as random forest, decision tree, K-nearest neighbors Naïve Bayes, and neural networks to enhance the cancer classification process.

The objectives of this paper are to propose a hybrid feature selection methods using the combination of filter and wrapper methods and apply them with different machine learning and ensemble learning methods to improve the performance of cancer classification.

The rest of the paper is structured as follows: Materials and Methods are provided in Section II. The experimental design is presented in Section III. Section IV shows the results and discussion. The conclusion and future work are presented in Section V.

II. Materials and Methods

A. Datasets

The proposed methods have been applied on four high dimensional microarray datasets for gene expression of different types of cancers. In addition to Breast Cancer and Brain Cancer dataset, Lung Cancer, Leukemia Cancer, Central Nervous System Cancer (CNS) datasets as shown in Table I. In the previous studies, other datasets have been used such as SRBCT, Prostate, Ovarian, MLL, Lymphoma, Leukemia and Colon, but the dimensionality of the genes for these methods is not too high and the applied feature selection and machine learning methods on these datasets obtained satisfactory performance.

TABLE I. Description of Datasets

Dataset	# Features	# Instances	# Classes
Brain	5597	42	5(10,10,10,4,8)
Breast	24481	97	2(46,51)
Lung	12600	203	5(139,17,6,21,20)
CNS	7129	60	2(21,39)

The Brain cancer [36] dataset includes 42 samples, with 5597 genes and five classes. The Breast dataset [37] includes 97 samples; with 24,481 genes. From these samples, 46 were classified as cancer. The Lung dataset [38] includes 203 samples with five classes. The number of features are 12,600 genes. Finally, the CNS dataset includes 60 samples, among these samples, only 21 are classified as cancer. The number of features are 7129 genes.

B. Hybrid Feature Selection Methods

In this study, several combinations between Filter-based and Wrapper-based feature selection methods have been done to suggest the better hybrid method. In Filter-based method, the information gain was used to reduce the dimensionality 1% and 5%. After that several wrapper-based methods were applied to investigate on the performance of gene selections, which are Best First, Greedy Stepwise, and Rank Search. Two classification methods were used in each wrapper method, which are: K-nearest neighbors and Naïve Bays. Fig. 1 shows the overall methods used in this study.

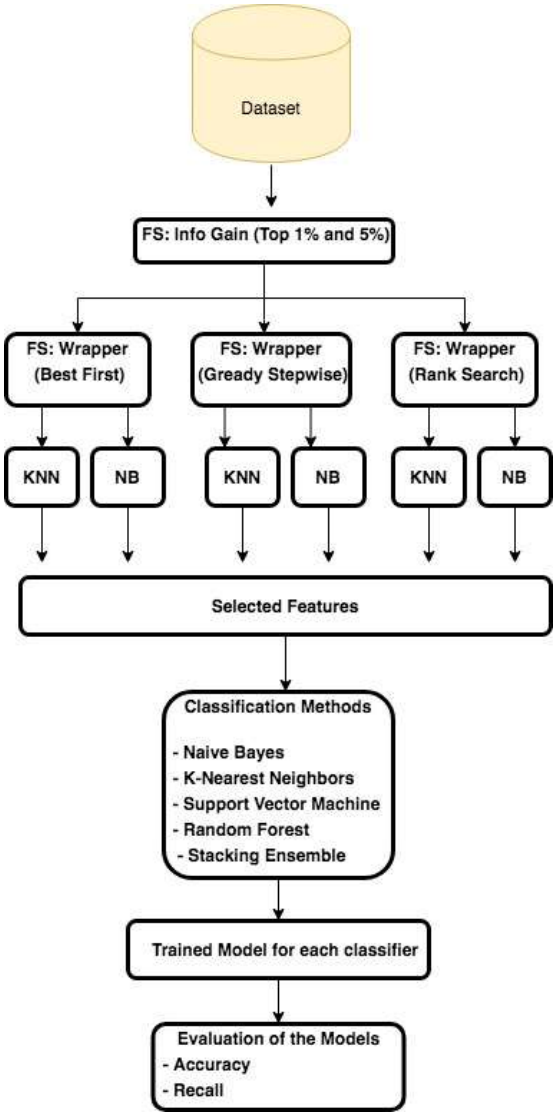


Fig. 1. The Developed Methods.

C. Machine Learning Methods

Several machine learning methods were applied for each combination in the feature selection step. These methods include individual and ensemble classification methods such as K-nearest neighbors, Naïve Bays, Support Vector Machine, Random Forests and Stacking Ensemble methods. The performance of these methods was evaluated before and after using the different combinations of feature selection and the best preforming methods were reported, as shown in Fig. 1.

III. Experimental Design

The experiments have been conducted on WEKA tool version 3.8. Each outcome of feature selection method has been fed to all machine learning methods (KNN, NB, SVM, RM and Stacking) in order to evaluate the performance of the gene selection and the cancer classification methods.

10-folds cross validation has been used for training and testing each dataset for all obtained combinations. The performance was evaluated using Accuracy and Recall measures, which are defined in the following equations (1) and (2).

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

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

where TP is true positive; TN is true negative; FP is false positive, and FN is false negative.

In addition, the performance of each method was compared before and after using features selection methods in order to discuss the enhancements obtained.

IV. Results and Discussion

The performance of the different combinations of feature selection and machine learning methods is shown in the tables below. The best performing method for each combination is bolded and the best performing method among all combinations for each dataset is shaded.

For Breast Cancer dataset, the performance of the used methods (using top 1% and 5% in the ranking method: information gain) are presented in Tables II and III.

As shown in Table II, the random forest method obtained the best accuracy and recall values with high dimensionality case (all features: 24481 and top 1% features: 244). However, after applying different combinations using ranking and wrapper methods, we found that Information Gain & Wrapper (NB & Best First) and Information Gain & Wrapper (NB & Gready Stepwise) obtained the best performance compared to all other methods/combinations before and after applying feature selection. Similarly, when the top 5% genes were selected in the ranking method, the performance of the used methods in Table III showed that random forest obtained the best results when high dimensional dataset was used, but when wrapper methods were applied, the combination of Information Gain and Wrapper (NB & Best First) obtained the best results. For Brain Cancer dataset, the results of used methods using the top 1% and 5% features are shown in Tables IV and V.

TABLE II. The Feature Selection and Machine Learning Methods for Breast Cancer Dataset using the Top 1% Features

	No. of Features	**Measure**	**NB**		**SVM**	**RF**	**Stacking**
All Features:	24481	Accuracy	0.546	0.608	0.546	**0.659**	0.526
		Recall	0.546	0.610	0.546	**0.660**	0.526
Information Gain 1 %:	244	Accuracy	0.608	0.804	0.722	**0.845**	0.814
		Recall	0.608	0.804	0.722	**0.845**	0.814
Info Gain & Wrapper (KNN & Best First)	9	Accuracy	0.577	**0.876**	0.629	0.814	0.763
		Recall	0.577	**0.876**	0.629	0.814	0.763
Info Gain & Wrapper (NB & Best First)	11	Accuracy	**0.938**	0.804	0.784	0.856	0.835
		Recall	**0.938**	0.804	0.784	0.856	0.835
Info Gain & Wrapper (KNN & GreadyStepwise)	5	Accuracy	0.557	**0.845**	0.639	0.825	0.763
		Recall	0.557	**0.845**	0.639	0.825	0.763
Info Gain & Wrapper (NB & GreadyStepwise)	11	Accuracy	**0.938**	0.804	0.784	0.856	0.835
		Recall	**0.938**	0.804	0.784	0.856	0.835
Info Gain & Wrapper (KNN & RankSearch)	104	Accuracy	0.577	**0.866**	0.732	0.835	0.794
		Recall	0.577	**0.866**	0.732	0.835	0.794
Info Gain & Wrapper (NB & RankSearch)	2	Accuracy	**0.742**	0.660	0.711	0.732	0.670
		Recall	**0.742**	0.660	0.711	0.732	0.670

TABLE III. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR BREAST CANCER DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	24481	Accuracy	0.546	0.608	0.546	**0.659**	0.526
		Recall	0.546	0.610	0.546	**0.660**	0.526
Information Gain 5 %:	1224	Accuracy	0.577	0.773	0.670	**0.814**	0.722
		Recall	0.577	0.773	0.670	**0.814**	0.722
Info Gain & Wrapper (KNN & Best First)	9	Accuracy	0.557	**0.907**	0.557	0.845	0.845
		Recall	0.557	**0.907**	0.557	0.845	0.845
Info Gain & Wrapper (NB & Best First)	15	Accuracy	**0.969**	0.784	0.825	0.866	0.928
		Recall	**0.969**	0.784	0.825	0.866	0.928
Info Gain & Wrapper (KNN & GreadyStepwise)	6	Accuracy	0.577	**0.897**	0.588	0.825	0.814
		Recall	0.577	**0.897**	0.588	0.825	0.814
Info Gain & Wrapper (NB & GreadyStepwise)	12	Accuracy	**0.949**	0.825	0.753	0.876	0.876
		Recall	**0.948**	0.825	0.753	0.876	0.876
Info Gain & Wrapper (KNN & RankSearch)	669	Accuracy	0.577	0.845	0.691	**0.866**	0.856
		Recall	0.577	0.845	0.691	**0.866**	0.856
Info Gain & Wrapper (NB & RankSearch)	2	Accuracy	**0.742**	0.660	0.711	0.732	0.670
		Recall	**0.742**	0.660	0.711	0.732	0.670

As shown in Tables IV and V, it is clearly shown that there are high improvements when using the combined feature selection methods. The best reported method is KNN as classification method and Information Gain & Wrapper (KNN & Best First) as feature selection methods using the top 1% and 5% features. In addition, for the top 5% features, other combinations obtained the same best results which are KNN classifier with Information Gain & Wrapper (KNN & GreadyStepwise), NB classifier with Information Gain & Wrapper (NB & Best First) and NB classifier with Information Gain & Wrapper (NB & GreadyStepwise) feature selection methods.

For Lung Cancer Dataset, the best performing method is NB classifier with Information Gain & Wrapper (NB & Best First) feature selection method for the top 1 % features (as shown in Table VI), and KNN with Info Gain & Wrapper (KNN & Best First) and Info Gain & Wrapper (KNN & GreadyStepwise) for the top 5% features (see Table VII).

TABLE IV. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR BRAIN CANCER DATASET USING THE TOP 1% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	5597	Accuracy	0.714	0.762	0.691	0.786	**0.881**
		Recall	0.714	0.762	0.690	0.786	**0.881**
Information Gain 1 %:	56	Accuracy	0.810	0.881	0.833	**0.905**	0.833
		Recall	0.810	0.881	0.833	**0.905**	0.833
Info Gain & Wrapper (KNN & Best First)	9	Accuracy	0.904	**0.100**	0.810	0.880	0.905
		Recall	0.905	**0.100**	0.810	0.881	0.905
Info Gain & Wrapper (NB & Best First)	11	Accuracy	**0.976**	0.881	0.786	0.952	0.857
		Recall	**0.976**	0.881	0.786	0.952	0.857
Info Gain & Wrapper (KNN & GreadyStepwise)	6	Accuracy	0.762	**0.952**	0.833	0.881	0.643
		Recall	0.762	**0.952**	0.833	0.881	0.643
Info Gain & Wrapper (NB & GreadyStepwise)	11	Accuracy	**0.976**	0.881	0.786	0.952	0.857
		Recall	**0.976**	0.881	0.786	0.952	0.857
Info Gain & Wrapper (KNN & RankSearch)	26	Accuracy	0.857	0.905	0.881	0.905	**0.929**
		Recall	0.857	0.905	0.881	0.905	**0.929**
Info Gain & Wrapper (NB & RankSearch)	9	Accuracy	0.881	0.857	0.857	0.881	**0.929**
		Recall	0.881	0.857	0.857	0.881	**0.929**

TABLE V. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR BRAIN CANCER DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	5597	Accuracy	0.714	0.762	0.691	0.786	**0.881**
		Recall	0.714	0.762	0.690	0.786	**0.881**
Information Gain 5 %:	280	Accuracy	0.810	0.857	**0.905**	0.881	0.810
		Recall	0.810	0.857	**0.905**	0.881	0.810
Info Gain & Wrapper (KNN & Best First)	11	Accuracy	0.905	**1.000**	0.810	0.881	0.810
		Recall	0.905	**1.000**	0.810	0.881	0.810
Info Gain & Wrapper (NB & Best First)	8	Accuracy	**1.000**	0.952	0.905	1.000	0.952
		Recall	**1.000**	0.952	0.905	1.000	0.952
Info Gain & Wrapper (KNN & GreadyStepwise)	9	Accuracy	0.786	**1.000**	0.810	0.833	0.810
		Recall	0.786	**1.000**	0.810	0.833	0.810
Info Gain & Wrapper (NB & GreadyStepwise)	8	Accuracy	**1.000**	0.952	0.905	**1.000**	0.952
		Recall	**1.000**	0.952	0.905	**1.000**	0.952
Info Gain & Wrapper (KNN & RankSearch)	191	Accuracy	0.833	0.905	0.929	**0.976**	0.905
		Recall	0.833	0.905	0.929	**0.976**	0.905
Info Gain & Wrapper (NB & RankSearch)	8	Accuracy	**0.929**	0.762	0.738	0.905	0.786
		Recall	**0.929**	0.762	0.738	0.905	0.786

Finally, the performance of the combined methods for CNS Dataset is presented in Tables VIII and IX. The results show that the best performing method is KNN classifier with Information Gain & Wrapper (KNN & Best First) and NB with Information Gain & Wrapper (NB & Best First) feature selection methods for the top 1 % features (as shown in Table VIII). In addition, the RF with the combination of Info Gain & Wrapper (KNN & RankSearch) obtained the same best results here. For the top 5% features, and KNN with Info Gain & Wrapper (KNN & Best First) consistently obtained the best results in this case as well.

By comparing the performances of all combined feature selection methods with different individual and ensemble machine learning methods, it is clearly shown that using these combinations with high dimensional datasets improved the cancer classification using all datasets used. The results in Tables II to IX showed that the best performing methods were KNN classifier with Information Gain & Wrapper (KNN & Best First) feature selection method and NB classifier with Info Gain & Wrapper (NB & Best First) feature selection method. Each one obtained the best five from eight cases using different datasets and different thresholds in the ranking methods (top 1% and 5% of features).

TABLE VI. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR LUNG CANCER DATASET USING THE TOP 1% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	12600	Accuracy	0.808	**0.897**	0.685	0.882	0.872
		Recall	0.808	**0.897**	0.685	0.882	0.872
Information Gain 1 %:	126	Accuracy	0.951	**0.956**	0.685	0.941	0.916
		Recall	0.951	**0.956**	0.685	0.941	0.916
Info Gain & Wrapper (KNN & Best First)	10	Accuracy	0.867	**0.970**	0.685	0.921	0.897
		Recall	0.867	**0.970**	0.685	0.921	0.897
Info Gain & Wrapper (NB & Best First)	15	Accuracy	**0.990**	0.902	0.685	0.951	0.946
		Recall	**0.990**	0.901	0.685	0.951	0.946
Info Gain & Wrapper (KNN & GreadyStepwise)	8	Accuracy	0.906	**0.966**	0.685	0.926	0.897
		Recall	0.906	**0.966**	0.685	0.926	0.897
Info Gain & Wrapper (NB & GreadyStepwise)	13	Accuracy	**0.985**	0.916	0.685	0.931	0.926
		Recall	**0.985**	0.916	0.685	0.931	0.926
Info Gain & Wrapper (KNN & RankSearch)	119	Accuracy	0.941	**0.966**	0.685	0.946	**0.966**
		Recall	0.941	**0.966**	0.685	0.946	**0.966**
Info Gain & Wrapper (NB & RankSearch)	126	Accuracy	0.951	**0.956**	0.685	0.941	0.916
		Recall	0.951	**0.956**	0.685	0.941	0.916

TABLE VII. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR LUNG CANCER DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	12600	Accuracy	0.808	**0.897**	0.685	0.882	0.872
		Recall	0.808	**0.897**	0.685	0.882	0.872
Information Gain 5 %:	630	Accuracy	0.941	**0.956**	0.685	0.941	**0.956**
		Recall	0.941	**0.956**	0.685	0.941	**0.956**
Info Gain & Wrapper (KNN & Best First)	11	Accuracy	0.892	**0.990**	0.685	0.936	0.926
		Recall	0.892	**0.990**	0.685	0.936	0.926
Info Gain & Wrapper (NB & Best First)	12	Accuracy	**0.985**	0.931	0.685	0.936	0.970
		Recall	**0.985**	0.931	0.685	0.936	0.970
Info Gain & Wrapper (KNN & GreadyStepwise)	11	Accuracy	0.892	**0.990**	0.685	0.936	0.921
		Recall	0.892	**0.990**	0.685	0.936	0.926
Info Gain & Wrapper (NB & GreadyStepwise)	12	Accuracy	**0.985**	0.931	0.685	0.936	0.970
		Recall	**0.985**	0.931	0.685	0.936	0.970
Info Gain & Wrapper (KNN & RankSearch)	213	Accuracy	0.936	**0.970**	0.685	0.936	0.961
		Recall	0.936	**0.970**	0.685	0.936	0.961
Info Gain & Wrapper (NB & RankSearch)	232	Accuracy	0.946	**0.961**	0.685	0.936	0.941
		Recall	0.946	**0.961**	0.685	0.936	0.941

TABLE VIII. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR CNS DATASET USING THE TOP 1% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	7129	Accuracy	0.617	0.567	0.650	**0.667**	0.550
		Recall	0.617	0.567	0.650	**0.667**	0.550
Information Gain 1 %:	71	Accuracy	0.717	0.817	0.650	**0.833**	0.767
		Recall	0.717	0.817	0.650	**0.833**	0.767
Info Gain & Wrapper (KNN & Best First)	7	Accuracy	0.733	**0.900**	0.650	0.833	0.867
		Recall	0.733	**0.900**	0.650	0.833	0.867
Info Gain & Wrapper (NB & Best First)	12	Accuracy	**0.900**	0.783	0.650	0.883	0.833
		Recall	**0.900**	0.783	0.650	0.883	0.833
Info Gain & Wrapper (KNN & GreadyStepwise)	3	Accuracy	0.600	**0.883**	0.650	0.750	0.767
		Recall	0.600	**0.883**	0.650	0.750	0.767
Info Gain & Wrapper (NB & GreadyStepwise)	6	Accuracy	**0.850**	0.583	0.650	0.800	0.700
		Recall	**0.850**	0.583	0.650	0.800	0.700
Info Gain & Wrapper (KNN & RankSearch)	40	Accuracy	0.767	0.883	0.650	**0.900**	0.817
		Recall	0.767	0.883	0.650	**0.900**	0.817
Info Gain & Wrapper (NB & RankSearch)	55	Accuracy	0.750	0.850	0.650	**0.867**	0.767
		Recall	0.750	0.850	0.650	**0.867**	0.767

TABLE IX. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR CNS DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	7129	Accuracy	0.617	0.567	0.650	**0.667**	0.550
		Recall	0.617	0.567	0.650	**0.667**	0.550
Information Gain 5 %:		Accuracy	0.667	0.617	0.650	**0.783**	0.733
		Recall	0.667	0.617	0.650	**0.783**	0.733
Info Gain & Wrapper (KNN & Best First)	13	Accuracy	0.583	**0.967**	0.650	0.767	0.800
		Recall	0.583	**0.967**	0.650	0.767	0.800
Info Gain & Wrapper (NB & Best First)	11	Accuracy	**0.883**	0.800	0.650	0.800	0.833
		Recall	**0.883**	0.800	0.650	0.800	0.833
Info Gain & Wrapper (KNN & GreadyStepwise)	2	Accuracy	0.467	**0.800**	0.650	0.717	0.700
		Recall	0.467	**0.800**	0.650	0.717	0.700
Info Gain & Wrapper (NB & GreadyStepwise)	11	Accuracy	**0.883**	0.800	0.650	0.800	0.833
		Recall	**0.883**	0.800	0.650	0.800	0.833
Info Gain & Wrapper (KNN & RankSearch)	37	Accuracy	0.750	0.850	0.650	**0.883**	0.750
		Recall	0.750	0.850	0.650	**0.883**	0.750
Info Gain & Wrapper (NB & RankSearch)	55	Accuracy	0.750	0.850	0.650	**0.867**	0.833
		Recall	0.750	0.850	0.650	**0.867**	0.833

V. CONCLUSION AND FUTURE WORK

The investigation of high dimensionality issue in microarray datasets has been conducted in this paper. Several combinations of ranking methods (using information gain with threshold of 1% and 5%) and wrapper methods (using KNN and NB with Best First, Greedy Stepwise, and Rank Search) were used to select the most important genes for microarray datasets. These datasets included Breast Cancer, Brain Cancer, Lung Cancer and CNS datasets. The experimental results showed the consistent good performance of applying all feature selection methods comparing with the case when all features were used (no feature selection methods). Among these used methods, the KNN with Information Gain & Wrapper (KNN & Best First) and NB with Info Gain & Wrapper (NB & Best First) obtained the best performance and overcame all other methods. Therefore, this study recommends to use one of these methods on high dimensionally microarray methods with the aim of obtaining better cancer classification accuracy. Future works will investigate other hybrid and intelligent feature selection methods for cancer classification using microarray datasets.

ACKNOWLEDGMENT

The authors would like to thank Deanship of Scientific Research at Al Imam Mohammad ibn Saud Islamic university, Saudi Arabia, for financing this project under the grant no. (18-11-09-015).

REFERENCES

[1] Y. Lu, and J. Han, "Cancer Classification Using Gene Expression Data," Information Systems, vol. 28, pp. 243-268, 2003.

[2] L.Yu, and H. Liu, "Redundancy based Feature Selection for Microarray Data," in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, pp. 737-742, 2004.

[3] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods", Information Sciences, 2014;282:111-35.

[4] G. Cosma, D. Brown, M. Archer, M. Khan, and A. G. Pockley, "A survey on computational intelligence approaches for predictive modeling in prostate cancer", Expert Systems with Applications, 70:1-19, 2017.

[5] R. K. Singh, and M. Sivabalakrishnan , "Feature selection of gene expression data for cancer classification: a review", Procedia Computer Science, 50:52-7, 2015.

[6] L.Wang, Feature selection in bioinformatics. 2012.

[7] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data", IEEE Transactions on Knowledge and Data Engineering, 25(1):1-14, 2013.

[8] P. C. Conilione, and D. Wang., "A comparative study on feature selection for E. coli promoter recognition", Intertional Journal of Information Technology,11(8):54-66, 2005.

[9] Z. M. Hira, and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data", Advances in bioinformatics, 2015.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data", Knowledge and Information Systems, 34(3):483-519, 2013.

[11] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, et al., "A survey on filter techniques for feature selection in gene expression microarray analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4):1106-19, 2012.

[12] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics", bioinformatics, 23(19):2507-17, 2007.

[13] M. A. Hall, Correlation-based feature selection for machine learning, 1999.

[14] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers", Genome Research, 11(11):1878-8.7, 2001.

[15] L. E. A. d. S. Santana, A. M. de Paula Canuto, "Filter-based optimization techniques for selection of feature subsets in ensemble systems", Expert Systems with Applications, 41(4):1622-31, 2014.

[16] P. M. Narendra, and K. Fukunaga, "A branch and bound algorithm for feature subset selection," IEEE Transactions on Computers, vol. C-26, no. 9, pp. 917-922, Sep. 1977.

[17] E. Alba, J. Garcia-Nieto, L. Jourdan, E-G Talbi, editors, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms", 2007 IEEE Congress on Evolutionary Computation, 2007.

[18] S. S. Hameed, R. Hassan, and F. F. Muhammad, "Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm", PLOS ONE, 12(11): 2017, e0187371. doi: 10.1371/journal.pone.0187371.

[19] L. Huijuan, C. Junying, Y. Ke, J. Qun, X. Yu, and G. Zhigang, "A hybrid feature selection algorithm for gene expression data classification", Neurocomputing, 256: 56-62, 2017.

[20] L-F Chen, C-T Su, K-H Chen, P-C Wang, "Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis", Neural Computing and Applications, 21(8):2087-96, 2012.

[21] T. Latkowski, and S. Osowski, "Data mining for feature selection in gene expression autism data", Expert Systems with Applications, 42(2):864-72, 2015.

[22] Y. Chen, D. Miao, and R. Wang, "A rough set approach to feature selection based on ant colony optimization" , Pattern Recognition Letters, 31(3):226-33, 2010.

[23] F. González, and L. A. Belanche, "Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy", arXiv preprint arXiv:13021733, 2013.

[24] F. Ardjani, K. Sadouni, and M. Benyettou, "Optimization of SVM multiclass by particle swarm (PSO-SVM)", 2nd IEEE International Workshop on Database Technology and Applications (DBTA), 2010.

[25] B. Tran, B. Xue, and M. Zhang, "Improved PSO for feature selection on high-dimensional datasets', Asia-Pacific Conference on Simulated Evolution and Learning, Springer ,2014.

[26] B. Ghaddar, and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines", European Journal of Operational Research, 265(3):993-1004, 2018.

[27] M. Dashtban, M. Balafar, and P. Suravajhala, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach", Genomics, 110(1):10-7, 2018.

[28] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification", Computational Biology and Chemistry, 56:49-60, 2015.

[29] E. Pashaei, and N. Aydin, "Binary black hole algorithm for feature selection and classification on biological data", Applied Soft Computing, 56:94-106, 2017.

[30] R. Aziz, C. K. Verma, M. Jha, and N. Srivastava, "Artificial neural network classification of microarray data using new hybrid gene selection method", International Journal of Data Mining and Bioinformatics, 17(1):42-65, 2017.

[31] S. A. Ludwig, S. Picek, and D. Jakobovic, "Classification of Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm", In: Kahraman C, Topcu YI, editors. Operations Research Applications in Health Care Management, Cham: Springer International Publishing, p. 327-47, 2018.

[32] C. S. R. Annavarapu, S. Dara, and H. Banka, "Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm", EXCLI journal,15:460, 2016.

[33] S. Kar, K. D. Sharma, M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique", Expert Systems with Applications, 42(1):612-27, 2015.

[34] I. Jain, V. K. Jain, andR. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification", Applied Soft Computing, 62:203-15, 2018.

[35] T. Almutiri, and F. Saeed, "Chi Square and Support Vector Machine with Recursive Feature Elimination for Gene Expression Data Classification", In 2019 IEEE First International Conference of Intelligent Computing and Engineering (ICOICE), December. (pp. 1-6), 2019.

[36] M. Dettling, and P. Bühlmann, "Supervised clustering of genes. Genome biology", 3(12), 1-15, 2002.

[37] Y. Sun, V. Urquidi, and S. Goodison, "Derivation of molecular signatures for breast cancer recurrence prediction using a two-way validation approach", Breast cancer research and treatment, 119(3), 593-599, 2010.

[38] http://www.biolab.si/supp/bi-cancer/projections/info/lung.html, last accessed 2010/05/20.

Visibility and Ethical Considerations of Pakistani Universities Researchers on Google Scholar

Muhammad Asghar Khan[1], Tariq Rahim Soomro[2]
College of Computer Science and Information Systems Institute of Business Management
Karachi, Pakistan

***Abstract*—Maximizing visibility by using academic profiling sites is very crucial in the academic world to improve the readership of research papers published and constant evaluation of research quality. In this article, the authors focused on the visibility of Pakistani University scholars on Google Scholar (GS). An intelligent Web Bot (MAKGBOT) was developed to collect the scholarly data of all Pakistani scholars, whose data is publicly available on Google Scholar. The findings of this research show that 87% of Pakistani universities have a presence on Google Scholar. It analyzes the research performance of scholars based on the last five years' data from 2016 to 2020. Furthermore, the analysis reports the level of scholarly activities of all provinces and autonomous areas of Pakistan. This paper concludes by discussing the ethical issue of misrepresentation of information on the public profile and its consequences on the rankings of legitimate scholars.**

***Keywords*—Google scholar; research visibility; Pakistani researchers; ethical considerations; web bot; research in Pakistan**

I. Introduction

The establishment of a research University is an expensive task compared to a typical educational institute. Research Universities are expensive because they need to attract good researchers and provide state-of-the-art infrastructure for teaching and research environment. In developed countries, governments and organizations are investing an enormous amount of resources in research and development. Multinational companies sponsor research-funded projects at Universities to get the solution to the real-life problem they are facing at a lower cost compared to the establishment of their research and development team. However, in developing countries, the establishment of a research University is even a more difficult task due to the scarcity of resources. Pakistan as a third-world country is also facing a shortage of research Universities. Regulatory bodies, such as, the Higher Education Commission (HEC) of Pakistan is working hard to improve the research culture in existing Universities of Pakistan.

Researchers produce research papers to highlight their contribution to the domain of their specialization. In the last few years, there is a great motivation to measure the quality of research of individuals based on different indicators [1]. Two most common indicators are the number of the paper published by an author and the number of citations [2]. There is an interesting discussion within bibliometrics that either researcher focuses on productivity or on the impact of the paper? [3] [4]. One of the major challenges for a researcher is to secure funding for a research project. Different higher education funding bodies provide sponsorship based on publication count (such as the Australian Funding System [3], whereas others focused on the quality of paper e.g. Netherlands national research assessment [4].

The scarcity of resources for research projects is a common issue, especially in developing economies. Funding organizations, such as, governments and R&D wings of multinational companies require to select competent researchers for their projects. The selection criteria are normally based on innovative idea and their impact on society after the completion of projects, however, the researcher's academic rankings and historical history of completion of projects were also considered when multiple competitors have the same level of creativity in their proposed projects. Different Indexing and abstracting services, such as, Scopus, Clarivative Analytics (Web of Science), and Google Scholar (GS) maintain the ranking of researches through different performance indicators, such as, the number of articles published, total citations, h-index, i10-index, Impact factor, etc. Google Scholar is a free and popular tool that helps to find scholars and their published articles along with performance indicators. Most University scholars create their Google Scholar profile to make their work visible over the Internet. In general, tight integration between Google Scholar and search engine improves the appearance of relevant search results from existing articles and helps to improve the number of a citation for authors of those papers.

This project extracts the important research matrices (for example, number of citations, the paper published, H-index, etc.) for all Pakistani university scholars whose profile is publicly available on GS. The authors of this study rank the universities based on their presence on GS and highlight the region of Pakistan where universities are more research-oriented. One of the contributions of this study is to list down the top researches of Pakistan and their academic affiliations. Furthermore, this study discusses the ethical issue of misrepresentation of information on academic profiles and its consequences on the ranking of legitimate scholars. The following sections provide a brief literature review, research methodology, limitation of the study, results & findings, ethical consideration, and finally discussion and future work.

II. Literature Review

Most of the scientific work nowadays is published in the form of research papers in journals or conferences, which can be easily found in bibliographic databases [5]. PubMed, ScienceDirect, Scopus, Web of Science, and Google Scholar are among the most famous bibliographic databases used to

find authors' profiles and relevant articles of interest by researchers. GS is a freely available academic search engine [6] that indexes scientific literature from a wide range of disciplines, record types, and languages, providing an outstanding set of additional offerings at the same time. The fact that it shows the number of citations obtained by each paper, irrespective of their origin, opened the door to a new type of bibliometric study, revolutionizing the comparison between academic performance, especially in the Humanities and Social Sciences [7]. Today, the majority of students and scholars are beginning to scan educational information in GS [8] [9]. Therefore, publications that are absent from the consequences pages of Google Scholar may also result in large readership losses and maybe even a decline in citations [10]. Anne-Wil Harzing in [11] claimed that GS can be used as a tool for citation analysis and described the benefits of GS over the ISI Web of Science along with the advantages and disadvantages of each tool.

Digital profiles are increasingly being used to assess potential writers, reviewers, and journal editors to exchange and collaborate on scholarly articles, and set up academic networks. Subsequently, simultaneous searches through the bibliographic databases and Websites, such as MEDLINE, Scopus, the Web of Science, and Google Scholar, make it possible to retrieve relevant items and navigate their extensive comparison through the author's profiles [12]. Editors of journals also refer to the profiles of their contributors in their editorial management systems, connected to bibliographic databases and search engines, in order to improve their quality and encourage the best contributors [13]. Furthermore, editors are strongly inspired to evaluate their contributors' academic profiles and Researcher IDs to avoid commenting on 'false' reviewers and misconduct of various types [14]. Alastair Smith [15] studied New Zealand's Performance-Based Research Funding (PBRF) evaluation system for universities and determined a very high correlation between PBRF output and the total number of citations return by GS. To improve the chance to secure more funding or to publish a paper in a reputed journal, few authors include fake papers in their profiles to increase the h-index and other research indicates or not verifying auto-generated papers suggested by Google Scholar, which results in more citations and number of publications. Such behavior raises many questions of ethical values, norms of societies, and financial pressures on researchers.

III. Methodology

A list of Pakistani Universities was retrieved from HEC [16]. HEC is the official body whose main responsibility is to regulate, fund, and accredited Universities and Degree Awarding Institutes (DAI) in Pakistan. There were 217 Universities/degree awarding institutes found in HEC accredited Universities database. HEC divided Pakistan into 4 provinces (Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan) and three autonomous areas (Gilgit Baltistan, Azad Jammu & Kashmir, and Islamabad Capital Territory).

The main aim of this study is to collect the data from Google Scholar profiles for all Pakistani University researchers, whose data is publicly available. The collection of scholarly data for the whole country is a difficult task if performed manually. Therefore, there is a need to automate this data collection process. Authors of the paper have developed a Web Bot called "MAKGBOT", which crawls all University scholars profile automatically and collect the following attributes from the publicly available profiles of the researchers on Google Scholar:

1) Google Scholar ID.
2) Google Scholar Name.
3) Total Citations.
4) Affiliation.
5) h-index.
6) i10-index.
7) Citations in the year 2016.
8) Citations in the year 2017.
9) Citations in the year 2018.
10) Citations in the year 2019.
11) Citations in the year 2020.
12) Citations in the last five year.
13) Total papers published by Scholar (complete).

MAKGBOT is a Python script, which uses the beautifulsoup [17] library to scrape information from the Google Scholar Website. Beautifulsoup is a very useful tool for searching, iterating, and modifying parse trees. MAKGBOT is similar to Scholarly [18], but it differs in a way that MAKGBOT can retrieve information for a University rather than a single author. Universities list along with URL can be fed to MAKGBOT as a comma-separated values (CSV) file instead of passing one University name at a time. Furthermore, MAKGBOT provides extra information, such as, total papers published by scholars who were not present in Scholarly.

IV. Limitations of the Study

The authors of this paper are aware that many Pakistani researchers have not created an account on Google Scholar or they have not set their profiles public. In this scenario, authors are unable to collect information about such researchers as the study focused was on the scholars whose profiles are publically available and visible. Furthermore, the authors noticed that a few authors have not changed their affiliation after switching their job to another institute. Therefore, the research contribution of such scholars will be counted towards their affiliation institute, which is verified, rather than where they are working currently. Furthermore, the authors are aware that the total number of papers and citations by a particular scholar may not be correct, as several authors set their profiles on the auto-generate mechanism and not annually verified the statistics and papers suggested by GS against their names.

Searching and collecting information for all researchers of a country is a time-consuming task. MAKGBOT restricted itself to limit the number of papers published by any author to 3000. As soon as a counter for the number of papers published by a single author reached 3000, MAKGBOT moved to the next scholar of that University on the list. Paper publication and the addition of new scholars on Google Scholar is a continuous process. Therefore, it is entirely possible that authors may miss a few scholars and papers, which were added

after 10 January 2021. There is a need for a system, which is capable to deals with continual queries and updates the existing records as soon as there are changes that exist on Google Scholar. Adding this capability in MAKGBOT is considered as future work.

V. Results and Findings

Based on the data selected by MAKGBOT until 10 January 2021, the following results are observed:

The pie chart in Fig. 1 shows the percentage share of scholars by provinces of Pakistan. Punjab is the largest province with the highest population and the maximum number of recognized Universities. Therefore, it is clear that the majority (33%) of the participants are from the Punjab region followed by Sindh (26%), Khyber Pakhtunkhwa (21%), Islamabad (12%), Balochistan (4%), Azad Jammu & Kashmir (3%) and Gilgit Baltistan (1%). Results from MAKGBOT illustrate that the visibility of researchers on Google Scholar is proportional to the population of those areas. The only exception is Islamabad capital territory, where the numbers of scholars on GS are relatively higher if compared to other autonomous areas of Pakistan. This variation is might due to the fact that Pakistan's best universities are located in Islamabad (HEC ranking [19], QS Ranking [20], Times University ranking [21]). Furthermore, Universities from the capital region have the highest Google Scholar visibility rate (100%) along with Gilgit Baltistan (100%), where only two Universities are situated.

Fig. 2 displays the top ten Pakistani Universities on Google Scholar based on the total number of scholars available and/or visible. Universities with the highest representation on Google Scholar are located either in Islamabad or Punjab. National University of Sciences and Technology ranked top with 764 active participants followed by the University of Lahore, University of Management & Technology, and Quaid e Azam University.

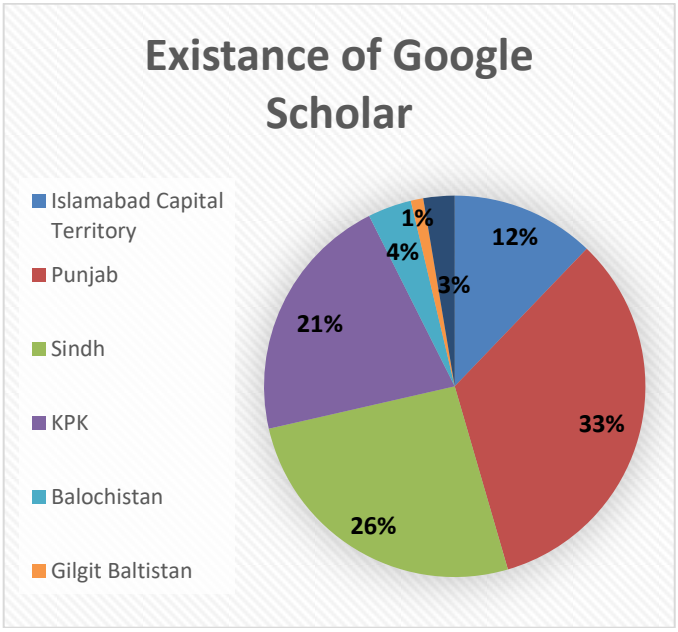


Fig. 1. The Percentage Share of Scholars by Provinces of Pakistan.

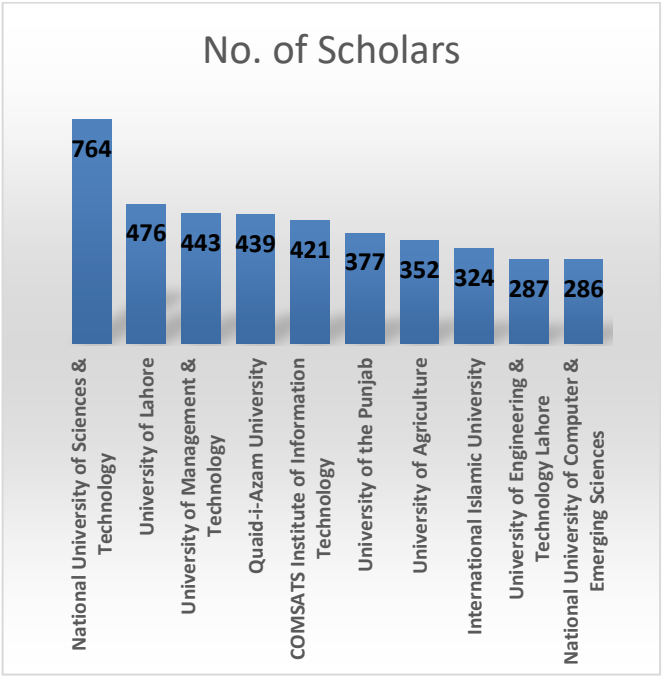


Fig. 2. Top Ten Pakistani Universities on Google Scholar by Number of Scholars.

The top three Universities in Punjab province by scholars count are the University of Lahore, University of Management and Technology, and the University of Punjab. The University of Lahore and the University of Management and Technology belong to the private sector. The University of Punjab is the oldest active University in Pakistan [22]. The top three Universities in Punjab are located in the provincial capital Lahore. There is a total of ten universities in the province of Punjab, which are not visible on GS. Table I shows the list of all Universities in Punjab province sorted by the total number of scholars visible.

The trend in Sindh province is slightly different as two out of three top positions are secured by public sector Universities. Aga Khan University ranked first, followed by Mehran University of Engineering and Technology and Univerity of Sindh, with scholar count 272, 255, and 244 respectively. Twelve Universities have no scholar whose profile is publically visible on Google Scholar. Table II illustrates the details of all Universities in the province of Sindh.

The visibility of scholars of Khyber Pakhtunkhwa Universities has resemblance with Sindh province as the top three positions were secured by public Universities. The number of scholar count difference between Abdul Wali Khan University and University of Engineering and Technology is only 23. Only one University of Khyber Pakhtunkhwa province is not visible at GS. Table III demonstrates the nitty-gritty rundown of all Universities in the region of Khyber Pakhtunkhwa.

Balochistan is the least populated province of Pakistan and only ten Universities are located in this region. A total of the three universities of Balochistan province are not visible on GS. Table IV highlights the important Google Scholar indicators.

TABLE I. LIST OF ALL OF THE UNIVERSITIES OF PUNJAB PROVINCE

University Name	Total Number of Scholars Visible	Max of h-index	Max of i10-index	Sum of total Citation	Sum of Total Paper Published
University of Lahore	476	92	354	123322	10322
University of Management & Technology	443	76	691	145506	17461
University of the Punjab	377	76	689	249661	23499
University of Agriculture	352	59	351	282540	23808
University of Engineering & Technology Lahore	287	81	780	141087	12379
Lahore University of Management Sciences	272	66	649	117894	9556
Government College for Women University	261	65	267	186651	15075
University of Sargodha	244	73	723	124734	12245
Bahauddin Zakariya University	239	71	638	213984	20075
University of Engineering & Technology Taxila	208	20	37	34790	4209
Islamia University	208	64	546	153240	14664
University of Gujrat	183	60	362	94814	9218
University of Veterinary & Animal Sciences	161	70	766	126843	14963
University of Central Punjab	160	39	123	28374	3851
Pir Mehr Ali Shah Arid Agriculture University	139	64	432	117638	9699
Government College University Lahore	127	45	138	62838	5533
University of Education	127	109	416	65497	4474
Lahore Garrison University	104	48	374	20068	3251
Information Technology University of the Punjab	95	33	75	17183	1519
Khawaja Freed University of Engineering & Information Technology	91	32	108	14860	2013
Lahore College for Women University	81	18	23	11384	1717
Forman Christian College	72	126	298	57448	2150
The Superior College	72	51	218	29610	2986
National Textile University	71	85	934	59762	7632
Muhammad Nawaz Shareef University of Agriculture	62	27	70	19868	3948
National University of Medical Sciences	52	51	477	39097	5086
Government College for Women University Sialkot	44	13	19	7499	859
University of Okara	40	14	20	5110	758
HITEC University	39	32	89	14197	1119
Fatima Jinnah Women University	38	16	37	6405	889
University of Wah	35	66	431	36505	3832
Minhaj University	30	14	23	2559	552
National College of Business Administration & Economics	29	15	19	2706	474
NFC Institute of Engineering & Technology	28	14	17	2318	359
King Edward Medical University	28	12	13	3540	670
University of Health Sciences	27	33	48	13823	1173
GIFT University	27	18	25	3850	360
Ghazi University	26	36	538	22951	3794
Kinnaird College for Women	26	13	18	2763	545
Cholistan University of Veterinary and Animal Sciences Bahawalpur	26	21	178	11073	2681
The University of Faisalabad	25	12	12	2361	606
Namal Institute Mainwali	24	17	32	6667	695
The Women University	23	13	16	2012	238
University of South Asia	21	73	625	31340	3125
University of Sialkot	18	35	104	11703	943

University of Sahiwal	16	16	19	2059	236
Institute of Southern Punjab	15	28	81	6360	791
Government Sadiq College Women University	12	10	10	1266	161
Lahore Leads University	10	6	4	363	78
Beaconhouse National University	9	29	50	6105	433
University of Narowal	6	8	5	369	64
Imperial College of Business Studies	5	16	19	1731	112
Muhammad Nawaz Sharif University of Engineering & Technology	5	24	61	2443	209
Rawalpindi Women University	4	4	1	126	29
Pakistan Institute of Fashion & Design	3	3	1	75	15
University of Home Economics Lahore	2	5	3	87	27
National College of Arts	2	10	12	390	55
Institute for Art and Culture	2	1	0	4	3
Institute of Management Sciences	2	32	106	4106	294
Nur International University	1	7	3	195	82
Fatima Jinnah Medical University Lahore	1	2	0	9	12
Qarshi University	1	1	0	4	1
University of Mainwali	1	18	30	1077	70
Ali Institute of Education	0	0	0	0	0
Hajvery University	0	0	0	0	0
Institute of Management Sciences	0	0	0	0	0
Lahore School of Economics	0	0	0	0	0
Punjab Tianjin University of Technology Lahore	0	0	0	0	0
Rawalpindi Medical University	0	0	0	0	0
Virtual University of Pakistan	0	0	0	0	0
Global Institute Lahore	0	0	0	0	0
Faisalabad Medical University Faisalabad	0	0	0	0	0
Times Institute Multan	0	0	0	0	0
Grand Total	**5615**	**126**	**934**	**2754844**	**267677**

TABLE II. LIST OF ALL OF THE UNIVERSITIES OF SINDH PROVINCE

University Name	**Total Number of Scholars Visible**	**Max of h-index**	**Max of i10-index**	**Sum of total Citation**	**Sum of Total Paper Published**
Aga Khan University	272	77	768	285556	22921
Mehran University of Engineering & Technology	255	27	54	46032	6319
University of Sindh	244	57	224	66177	7736
University of Karachi	227	76	769	208561	22388
Sukkur Institute of Business Administration	166	26	67	23311	4763
NED University of Engineering & Technology	141	23	36	25188	2907
Institute of Business Management	137	17	28	13298	2207
DOW University of Health Sciences	113	44	224	44058	4687
Quaid-e-Awam University of Engineering Sciences & Technology	96	14	16	11771	1747
Institute of Business Administration	73	11	11	5581	894
Iqra University	68	29	70	10060	1212
Shah Abdul Latif University	68	15	21	8694	1531
Sindh Agriculture University Tandojam	66	16	29	15273	2657
Shaheed Zulfikar Ali Bhutto Institute of Science & Technology	64	16	23	5669	790
Hamdard University	37	12	15	2601	692

Sir Syed University of Engineering & Technology	37	18	31	5020	821
Liaquat University of Medical & Health Sciences	37	15	30	7254	1470
Karachi Institute of Economics & Technology	36	12	13	4447	734
Sindh Madresatul Islam University	34	16	22	3413	498
Zia-ud-Din University	34	62	448	26485	3606
Dawood University of Engineering & Technology	27	21	29	5488	457
Jinnah Sindh Medical University	25	34	65	8891	822
Isra University	25	31	131	10317	1307
Indus University	24	14	18	1230	227
Shaheed Benazir Bhutto University Shaheed Benazirabad	24	7	6	919	279
Habib University	22	15	21	3465	407
Mohammad Ali Jinnah University	20	5	3	309	103
Barret Hodgson University	19	8	8	1604	265
Jinnah University for Women	16	10	12	1708	425
Benazir Bhutto Shaheed University Lyari	15	25	69	2895	297
KASB Institute of Technology	13	11	15	1053	158
Baqai Medical University	10	28	51	6574	593
Preston University	10	20	36	3107	317
Peoples University of Medical & Health Sciences for Women	9	10	14	869	212
Begum Nusrat Bhutto Women University Sukkur	6	15	21	805	88
Shaheed Mohtarma Benazir Bhutto Medical University	5	15	19	1343	73
Sindh Institute of Medical Sciences	4	25	74	3743	496
Karachi School for Business & Leadership	4	17	24	1957	121
Benazir Bhutto Shaheed University of Technology & Skill Development Khairpur Mirs	4	7	6	330	14
Shaheed Benazir Bhutto University of Veterinary & Animal Sciences	3	11	13	1924	72
University of Sufism and Modern Sciences Bhitshah	3	2	0	15	34
ILMA University	2	3	1	45	32
Newport Institute of Communications & Economics	1	1	0	1	16
Nazeer Hussain University	1	1	0	2	7
Shaheed Zulfiqar Ali Bhutto University of Law	1	5	1	154	61
Government College University Hyderabad	1	7	5	155	18
Textile Institute of Pakistan	1	2	1	17	8
Dadabhoy Institute of Higher Education	1	8	7	556	41
Emaan Institute of Management & Sciences	1	4	0	35	23
Gambat Institute of Medical Sciences	0	0	0	0	0
Greenwich University	0	0	0	0	0
Indus Valley School of Art & Architecture	0	0	0	0	0
Pakistan Naval Academy	0	0	0	0	0
Preston Institute of Management Science & Technology	0	0	0	0	0
Shaheed Benazir Bhutto City University	0	0	0	0	0
Shaheed Benazir Bhutto Dewan University	0	0	0	0	0
Shaheed Benazir Bhutto University	0	0	0	0	0
Sindh Institute of Management & Technology	0	0	0	0	0
Commecs Institute of Business & Emerging Sciences	0	0	0	0	0
The Shaikh Ayaz University Shikarpur	0	0	0	0	0
Sohail University Karachi	0	0	0	0	0
Grand Total	**2502**	**77**	**769**	**877960**	**97553**

TABLE III. LIST OF ALL OF THE UNIVERSITIES OF KHYBER PAKHTUNKHWA PROVINCE

University Name	Total Number of Scholars Visible	Max of h-index	Max of i10-index	Sum of total Citation	Sum of Total Paper Published
Abdul Wali Khan University	203	65	224	125957	12436
University of Engineering & Technology	180	42	209	42800	4771
The University of Agriculture Peshawar	167	69	668	242562	22600
University of Malakand	101	28	71	48273	4766
Hazara University	97	69	548	69170	7201
Ghulam Ishaq Khan Institute of Engineering Sciences & Technology	96	29	96	23788	2510
Kohat University of Science and Technology	91	51	500	56387	6122
University of Haripur	66	77	282	40119	2331
Gomal University	59	66	535	72979	8813
University of Peshawar	50	74	725	60776	5118
Islamia College University	48	26	64	19194	1506
Sarhad University of Science & Information Technology	47	18	42	8955	1287
University of Swat	46	26	39	12282	1015
University of Swabi	46	47	153	20023	1721
Institute of Management Sciences	42	61	545	29263	3921
Khyber Medical University	41	15	23	10418	1245
Shaheed Benazir Bhutto University	38	47	223	24598	2796
Abasyn University	33	27	56	7647	949
Bacha Khan University	31	53	484	23242	3508
CECOS University of Information Technology & Emerging Sciences	28	34	102	15183	966
Qurtaba University of Science & Information Technology	21	22	55	4089	255
University of Science & Technology	19	24	34	8164	450
Iqra National University	15	13	19	1291	336
Khushal Khan Khattak University	15	15	23	3066	552
University of Engineering & Technology (UET) Mardan	15	14	16	2170	546
Shaheed Benazir Bhutto Women University	14	9	8	1234	266
Abbottabad University of Science and Technology (AUST)	14	26	52	6448	561
Pak-Austria Fachhochschule Institute of Applied Sciences and Technology Haripur	12	20	41	4452	345
Northern University	10	65	564	26231	2947
Preston University	10	20	36	3107	317
Women University Swabi	8	18	28	2354	166
University of Chitral	8	10	10	327	66
University of FATA	7	12	12	1229	110
The University of Lakki Marwat	5	10	11	585	96
Shuhada-e-Army Public School University of Technology Nowshera	4	11	12	734	97
University of Buner	2	12	15	1155	41
Women University Mardan	1	2	0	14	3
City University of Science and Information Technology	1	1	0	3	3
Brains Institute Peshawar	1	1	0	4	3
Gandhara University	1	4	0	26	14
Pakistan Military Academy	0	0	0	0	0
Grand Total	**1693**	**77**	**725**	**1020299**	**102756**

TABLE IV. LIST OF ALL OF THE UNIVERSITIES OF BALUCHISTAN PROVINCE

University Name	Total Number of Scholars Visible	Max of h-index	Max of i10-index	Sum of total Citation	Sum of Total Paper Published
Balochistan University of Information Technology Engineering & Management Sciences (BUITEMS)	115	21	41	18383	2185
University of Balochistan	40	35	120	12518	1910
Lasbela University of Agriculture Water & Marine Sciences	19	21	40	3561	611
Balochistan University of Engineering & Technology	11	8	8	603	135
University of Turbat	7	11	14	1386	120
University of Loralai	6	6	4	740	78
Sardar Bahadur Khan Women University	3	4	1	100	33
Al-Hamd Islamic University	0	0	0	0	0
Mir Chakar Khan Rind University Sibi	0	0	0	0	0
The Bolan University of Medical & Health Sciences	0	0	0	0	0
Grand Total	**201**	**35**	**120**	**37291**	**5072**

Islamabad is the capital of Pakistan and most of the top-ranked universities of Pakistan are located in the Capital Territory. By and large the public visibility of researchers on Google Scholar is higher whenever contrasted with different areas of Pakistan and all 23 universities have visibility on Google Scholar. Table V shows the quick summary of all Universities in Islamabad.

Azad Jammu & Kashmir is an autonomous region of Pakistan. Mirpur University of Science & Technology and the University of Azad Jammu & Kashmir are two major Universities in this region. Two universities of Azad Jammu & Kashmir area are not visible on GS. Table VI features significant Google Scholar pointers of this state.

Gilgit Baltistan is a remote area of Pakistan with a limited population. Just two Universities are situated in this locale. Table VII shows the number of academicians publically visible on Google Scholar.

One of the main purposes of this study to identify the best researchers in Pakistan. Table VIII shows the list of top ten Pakistani Researchers based on their total number of citations.

The h-index is one of the main indicators to reflect the quality of research papers. Table IX displays a list of the top ten Pakistani researchers based on the Google Scholar h-index.

TABLE V. LIST OF ALL OF THE UNIVERSITIES OF ISLAMABAD CAPITAL TERRITORY

University Name	Total Number of Scholars Visible	Max of h-index	Max of i10-index	Sum of total Citation	Sum of Total Paper Published
National University of Sciences & Technology	764	54	239	202514	20809
Quaid-i-Azam University	439	79	695	423528	36318
COMSATS Institute of Information Technology	421	69	688	282658	19610
International Islamic University	324	81	822	207479	17095
National University of Computer & Emerging Sciences	286	45	198	57590	6449
Bahria University	200	29	94	30839	4658
Pakistan Institute of Engineering & Applied Sciences	157	75	815	110381	10823
Riphah International University	150	74	613	65224	7783
Air University	134	91	711	77741	5177
Shifa Tameer-e-Millat University	129	26	99	17318	2276
Foundation University Islamabad	97	36	69	19577	1735
National University of Modern Languages	76	17	34	6350	1082
Institute of Space Technology	65	34	66	25886	2026
Capital University of Science & Technology	49	22	60	15760	1468
Allama Iqbal Open University	45	24	47	11672	1172
National University of Technology (NUTECH) Islamabad	31	37	134	10563	937
Pakistan Institute of Development Economics (PIDE)	22	23	46	12498	729

Health Services Academy HSA Islamabad	14	53	74	43911	710
National Defense University	12	15	21	1572	358
Shaheed Zulfiqar Ali Bhutto Medical University	9	22	43	3778	892
Federal Urdu University of Arts Sciences & Technology	8	11	15	1328	191
Sir Syed (CASE) Institute of Technology	4	10	11	893	127
Muslim Youth University	4	2	0	11	7
Grand Total	**3440**	**91**	**822**	**1629071**	**142432**

TABLE VI. LIST OF ALL OF THE UNIVERSITIES OF AZAD JAMMU AND KASHMIR REGION

University Name	**Total Number of Scholars Visible**	**Max of h-index**	**Max of i10-index**	**Sum of total Citation**	**Sum of Total Paper Published**
Mirpur University of Science & Technology	91	34	151	25648	2570
University of Azad Jammu & Kashmir	71	75	669	98498	5329
University of Poonch	47	23	40	13217	3245
University of Kotli Azad Jammu and Kashmir	21	8	6	1357	202
Women University of Azad Jammu & Kashmir	9	10	11	1390	257
AlKhair University	0	0	0	0	0
Mohi-ud-Din Islamic University	0	0	0	0	0
Grand Total	**239**	**75**	**669**	**140110**	**11603**

TABLE VII. LIST OF ALL OF THE UNIVERSITIES OF GILGIT BALTISTAN

University Name	**Total Number of Scholars Visible**	**Max of h-index**	**Max of i10-index**	**Sum of total Citation**	**Sum of Total Paper Published**
Karakurum International University	68	78	770	58333	6253
University of Baltistan Skardu	11	13	13	2645	197
Grand Total	**79**	**78**	**770**	**60978**	**6450**

TABLE VIII. LIST OF TOP TEN PAKISTANI RESEARCHERS BASED ON TOTAL CITATIONS

Google Scholar ID	**Scholar Name**	**Affiliation**	**Total Citations**	**h-index**	**i10-index**	**Citations in last 5 Years**	**Total Paper Published**
vUSWHc8AAAAJ	Dr. Muhammad Naeem Ahmed	University of Azad Jammu & Kashmir	77300	75	669	27181	2904
_3WBQxYAAAAJ	Muhammad Arshad Sajjad	Air University	53397	91	711	41529	2002
B6TB8IEAAAAJ	Prof. Dr. Hidayatur Rahman	The University of Agriculture Peshawar	47277	69	668	28537	2978
ByAexSYAAAAJ	Muhammad Akbar Zafar Khan	Islamia University	41810	60	546	31374	2799
nAFs720AAAAJ	Aysha Habib Khan	Aga Khan University	41479	77	768	20955	2980
E82kqSgAAAAJ	Ejaz Khan	Health Services Academy HSA Islamabad	41218	53	74	40108	130
upXMs64AAAAJ	Prof. Dr. Muhammad Tahir Hussain	National Textile University	39567	85	934	20767	2979
Vqh3MKMAAAAJ	Dr. Farooq Ahmad	University of Engineering & Technology Lahore	38209	81	780	23245	2957
Hy-zuEwAAAAJ	SAIF UR REHMAN	University of Management & Technology	37991	76	691	19400	2976
-wnLx6gAAAAJ	Dr. Tania Ahmed Shakoori	University of Lahore	37938	92	354	18399	947

TABLE IX. LIST OF TOP TEN PAKISTANI RESEARCHERS BASED ON H-INDEX

Google Scholar ID	Scholar Name	Affiliation	h-Index	Total Citations	i10-index	Total Paper Published
9l8oSH0AAAAJ	Kauser Abdulla Malik	Forman Christian College	126	37925	298	622
QjPoerMAAAAJ	MA Saeed	University of Education	109	37220	416	1030
-wnLx6gAAAAJ	Dr. Tania Ahmed Shakoori	University of Lahore	92	37938	354	947
_3WBQxYAAAAJ	Muhammad Arshad Sajjad	Air University	91	53397	711	2002
upXMs64AAAAJ	Prof. Dr. Muhammad Tahir Hussain	National Textile University	85	39567	934	2979
Vqh3MKMAAAAJ	Dr. Farooq Ahmad	University of Engineering & Technology Lahore	81	38209	780	2957
KqaU3UMAAAAJ	Asma Hussain	International Islamic University	81	34537	822	2977
b_VQd2EAAAAJ	Amjad Khan	Quaid-i-Azam University	79	33803	586	2889
TqhlyQMAAAAJ	Asif Khan	Karakurum International University	78	36430	770	2985
nAFs720AAAAJ	Aysha Habib Khan	Aga Khan University	77	41479	768	2980

VI. ETHICAL CONSIDERATIONS

Ethical consideration is one of the most important parts of any kind of research. According to Bryman and Bell [23], it is mandatory for the author of a research paper to acknowledge the works of other researchers by use of the referencing system recommended by the publication committee of the journal, where the paper is supposed to be published. It is highly unethical if a scholar claimed the authorship of a paper, which is not written by him/her. In the previous section, Tables VIII and IX highlighted the top researchers of Pakistan based on total number of citations and highest h-index rankings respectively, which are currently updated on GS. The authors of this paper believe that many of the scholar's names shown in these tables did not verify their paper lists on GS, which is causing misrepresentation of profiles. Furthermore, such actions restraining other legitimate researchers to become visible on top of the list. There might be many reasons why scholars on GS are not validating their paper on GS. Two common reasons are mentioned in the remainder of this section.

1) GS automatically generate the list of citations and paper published by a scholar based on its ranking algorithms. Scholars are either to set manual update and verify each entry before being added to his/her profile, or scholars are supposed to deleted papers that are not written by them and mistakenly added to their profile due to similarity of name or co-author affiliations. However, due to time constraints and busy schedules, scholars are not visiting GS to verify their profiles regularly.

2) Scholars might deliberately add a few high-quality papers with higher citations in their profiles to improve their visibility on GS. According to the GS ranking algorithm [24], profiles with higher citations appear first in the university's GS list, as well as, on the Google search engine. It is a high probability that papers that appear in the top position might get more citations compared to new papers that get less attention from the visitors of GS as these papers hide at the bottom of the list. Other possible motives for adding non-legitimate papers are to impress peers or to gain research funding as most of the sponsors are looking for GS research indicators to select the best researchers for their projects.

VII. DISCUSSION AND FUTURE WORK

The Google scholar is a very popular and useful tool to showcase the author's profile over the internet. Many universities are using Google Suite for email and other administrative tasks, therefore, it is easier for them to integrate university faculty research profiles with GS. In this study, the authors collected the GS scholar data (total 13769 Scholars) of all 217 Pakistani recognized universities. Twenty-eight universities have no representation on Google Scholar at all. Results showed that universities from Islamabad Capital Territory have high visibility compared to other autonomous areas of Pakistan. In general, the number of scholars' visibility on GS is logical and it is representing the population of the four provinces of Pakistan, where, Punjab is leading followed by Sindh, Khyber Pakhtunkhwa, and Baluchistan. However, individual academic indicators of many top researchers of Pakistan from their public profiles are misrepresenting and they contained papers and citations, which may not belong to specific scholars.

Misrepresentation of information on public profiles is a serious ethical issue. This misrepresentation of data might be the result of the auto-generation of citation by GS or any other social stress on the scholar by the academic environment. It is the responsibility of the scholars to make sure that they frequently check their GS profiles and remove papers that were added by GS automatically in authors' profiles, which were not written by them. As a responsible citizen of the research community, scholars should only take ownership of those papers on public profiles that were produced by them and not those whose authors' names are similar to them. Furthermore, there is a need for strict control on GS, which makes sure only legitimate papers are added on GS public profiles, both by scholars and auto-recommendation features of GS. The GS may add the feature of verifying from one of the co-authors the legitimation ownership of the scholar.

Research plays a vital role in the ranking of universities. Universities evaluating bodies that issue university rankings, such as, Higher Education Commission of Pakistan, QS, and Time university rankings required universities to provide them

their research outcomes on a yearly basis. GS is a very useful tool for universities to publicly present their research achievement systematically. As a regulatory body, HEC has a good influence over Pakistan universities. Therefore, it is easier for HEC to advise universities to maintain their research activities on GS. Universities can make sure that all scholars affiliated with their university have legitimate public profiles visible on GS and scholars update their profiles regularly to avoid ethical and social issues, which were discussed in the previous section. Universities/DAI make sure that scholars will not get any benefit because of his/her incorrect GS profile.

Limitations of this study are already discussed in a dedicated section. The authors are planning to add two more features in MAKGBOT to overcome these limitations. Firstly, MAKGBOT should support real-time or periodic updates. As soon as, the new university is added to the HEC repository or the paper is published in the public profile of a Pakistani scholar, the system will update the results and provide basic reports. Secondly, MAKGBOT should check the legitimacy of the papers by verifying the scholar's name in the published paper, which is added to the public profile of GS.

REFERENCES

[1] U. Sandstrom and P. v. d. Besselaar, "Quantity and/or Quality? The Importance of Publishing Many Papers," PLoS ONE, vol. 11, no. 11, pp. 1-16, 2016.

[2] C. Bosquet and P.-P. Combes, "Are academics who publish more also more cited? Individual determinants of publication and citation records," Scientometrics,Springer,Akadémiai Kiadó, vol. 97, no. 3, pp. 831-857, 2013.

[3] L. Butler, "Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts," Research Policy, vol. 32, no. 1, pp. 143-155, 2003.

[4] H. Dijstelbloem, F. Huisman, F. Miedema and W. Mijnhardt, "Science in Transition status report: Debate, progress and recommendations," Science in Transition, Amsterdam, 2014.

[5] M. Tober, "PubMed, ScienceDirect, Scopus or Google Scholar – Which is the best search engine for an effective literature research in laser medicine?," Medical Laser Application, vol. 26, no. 3, pp. 139-144, 2011.

[6] J. L. Ortega, "6 - Google Scholar: on the shoulders of a giant," in Academic Search Engines, Chandos Publishing, 2014, pp. 109-141.

[7] E. Orduna-Malea, J. M. Ayllón, A. Martín-Martín and E. D. López-Cózar , "Methods for estimating the size of Google Scholar," Scientometrics, vol. 104, no. 3, pp. 931-949, 2015.

[8] R. Housewright, R. C. Schonfeld and K. Wulfson, "Ithaka S+R|JISC|RLUK UK Survey of Academics 2012," 2013.

[9] B. Kramer and J. Bosman, "Innovations in scholarly communication - global survey on research tool usage," F1000Research, vol. 5, no. 692, pp. 1-13, 2016.

[10] E. D. López-Cózar, E. Orduna-Malea, A. Martín-Martín and J. . M. Ayllón, "Google Scholar: The Big Data Bibliographic Tool," in Research Analytics: Boosting University Productivity and Competitiveness through Scientometrics, F. J. Cantú-Ortiz, Ed., CRC Press, 2017, pp. 59-80.

[11] A.-W. K. Harzing and R. v. d. Wal, "Google Scholar as a new source for citation analysis?," Ethics in Science and Environmental Politics, vol. 8, no. 1, pp. 61-73, 2008.

[12] S. Pylarinou and S. Kapidakis, "Tracking Scholarly Publishing of Hospitals Using MEDLINE, Scopus, WoS and Google Scholar," Journal of Hospital Librarianship, vol. 17, no. 3, pp. 209-216, 2017.

[13] A. Y. Gasparyan, B. Nurmashev, M. Yessirkepov, D. A. Endovitskiy, A. A. Voronov and G. D. Kitas, "Researcher and Author Profiles: Opportunities, Advantages, and Limitations," Journal of Korean medical science, vol. 32, no. 11, pp. 1749-1756, 2017.

[14] J. Gao and T. Zhou, "Retractions: stamp out fake peer review," Nature, vol. 546, p. 33, 2017.

[15] A. G. Smith, "Benchmarking Google Scholar with the New Zealand PBRF research assessment exercise," Scientometrics, vol. 74, no. 2, pp. 309-316, 2007.

[16] "HEC Recognised Universities and Degree Awarding Institutions," HEC, [Online]. Available: https://hec.gov.pk/english/Universities/pages/recognised.aspx. [Accessed 10 Jan 2021].

[17] "pypi.org," [Online]. Available: https://pypi.org/project/beautifulsoup4/. [Accessed 8 November 2020].

[18] "pypi.org," [Online]. Available: https://pypi.org/project/scholarly/. [Accessed 8 November 2020].

[19] "Universities Ranking," HEC, [Online]. Available: https://www.hec.gov.pk/english/universities/Pages/AJK/rank.aspx. [Accessed 26 Dec 2020].

[20] "QS World University Ranking," QS, [Online]. Available: https://www.topuniversities.com/university-rankings/world-university-rankings/2020. [Accessed 28 December 2020].

[21] "THE World University Rankings," THE, [Online]. Available: https://www.timeshighereducation.com/world-university-rankings/2020/world-ranking#!/page/0/length/25/locations/PK/sort_by/rank/sort_order/asc/cols/stats. [Accessed 28 December 2020].

[22] "University of Punjab," [Online]. Available: http://pu.edu.pk/page/show/historyandpride.html . [Accessed 28 December 2020].

[23] E. Bell, A. Bryman and B. Harley, Business Research Methods, 5th Edition ed., Oxford University Press, 2018.

[24] J. Beel and B. Gipp, "Google Scholar's Ranking Algorithm: An Introductory Overview," in Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), Rio de Janeiro (Brazil), 2009.

Early Detection of Severe Flu Outbreaks using Contextual Word Embeddings

Redouane Karsi[1], Mounia Zaim[2], Jamila El Alami[3]
LASTIMI Laboratory, Higher School of Technology of Sale
Mohammed V University
Rabat, Morocco

***Abstract*—The purpose of automated health surveillance systems is to predict the emergence of a disease. In most cases, these systems use a text categorization model to classify any clinical text into a category corresponding to an illness. The problem arises when the target classes refer to diseases sharing multiple information such as symptoms. Thus, the classifier will have difficulty discriminating the disease under surveillance from other conditions of the same family, causing an increase in misclassification rate. Clinical texts contain keywords carrying relevant information to distinguish diseases with similar symptoms. However, these specific words are rare and sparse. Therefore, they have a minor impact on machine learning models' performance. Assuming that emphasizing specific terms contributes to improving classification performance, we propose an algorithm that enriches training samples with terms semantically similar to specific terms using the deep contextualized word embeddings ELMo. Next, we devise a weighting scheme combining chi-square and semantic scores to reflect the relatedness between features and the disease under surveillance. We evaluate our model using the SVM algorithm trained on i2b2 dataset supplemented by documents collected from Ibn Sina hospital in Rabat. Experimental results show a clear improvement in classification performance than baseline methods with an F-measure reaching 86.54%.**

Keywords*—*ELMo; SVM; contextual word embeddings; semantic term weighting; health surveillance; text classification

I. Introduction

Public health surveillance is a significant focus of National health policies. It is ensured by collecting epidemiological data from various healthcare facilities to detect disease outbreaks and subsequently plan appropriate response strategies early.

In Morocco's epidemiological surveillance system, the law requires healthcare producers to report all confirmed cases of notifiable diseases. For this, physicians must fill a particular form with the patient's clinical and demographic data. However, many physicians do not respect this notification formality, especially in the private sector. Thus, the total amount of collected forms is not entirely significant for correctly estimating epidemiological trends. Besides, the notification procedures for disease and collected data processing are not automated for acquiring relevant epidemiological indicators in real-time.

Today, several hospitals in the country have implemented the electronic health record (EHR), which appears to be an excellent opportunity for a better epidemiological surveillance system, because patient information captured and stored in EHR are so relevant for healthcare decision making [1], [2].

In EHRs, data is captured in a structured format, such as administrative data. Simultaneously, there is unstructured data written in a free text by practitioners. This textual data reflects the patient's health status and helps determine exciting health indicators [3].

Text classification algorithms select meaningful information from EHRs to organize textual documents into a set of pre-defined categories [4]. In outbreak detection, a class corresponds to one disease.

Feature selection is a crucial element in the preprocessing phase. Its role is to optimally reduce feature space's dimensionality by selecting a subset of relevant terms according to some criteria [5].

The biggest challenge of feature selection methods is to correctly select features with high discriminative power. For this purpose, some methods rely on Frequency-based feature selection[6], [7], others like Information Gain (IG) and chi-square test rank terms according to their correlation with the class variable [8], [9]. More recently, a new research trend favors semantic similarity based on knowledge resources and the fast-growing field of deep neural networks [10].

The flu surveillance system's goal is to predict the spread of a severe form of influenza accurately. The acquired flu-related free-text clinical records are classified into two categories (severe flu or mild flu). These two forms of flu share many signs and symptoms. In this situation, the risk of misclassification increases, especially for documents related to severe influenza cases, since the frequency of specific features that characterize severe cases is low compared to common features frequency.

In this respect, many research efforts attempt to improve feature selection algorithms by highlighting the discriminative power of infrequent specific terms. Thus, ontology-based feature selection methods like UMLS and SNOMED CT have been intensively experimented with real improvements in the medical domain. [11], [12].

Despite the progress achieved in utilizing ontology-based feature selection methods, it is not sure that they are useful in differentiating between two similar classes as long as they share many terms. We can mention two reasons:

1) The clinical note: "The patient with fever, cough, and runny nose, diagnosed as positive for H1N1", reveals a severe case of flu, yet the three common terms in the document (fever, cough, and runny nose) are preponderant compared to the only specific term (H1N1) which despite its importance, it is very infrequent in the corpora, which might reduce classifiers efficiency [13].

2) We usually calculate feature weights according to their frequencies or their statistical correlation with the target class. However, the rare term "H1N1" can be underestimated despite being semantically more heavily weighted than all other features.

To overcome the shortcomings of statistical and ontology-based feature selection methods, static word embeddings models have been put forward because of their ability to capture word semantic proprieties [14]. However, they are inefficient in handling the widely varying medical spelling since they provide a unique word representation. Hence, we hypothesize that a contextual word embeddings representation [15] with a weighting scheme combining statistical and semantic scores can better emphasize rare medical words improving classification performance in outbreaks detection.

In this paper, we propose in a first step an algorithm that aims to enrich training samples related to severe cases of flu with features that are semantically similar to specific features. The idea behind this algorithm is to mitigate the deficiency caused by the scarcity of specific features by adding new features to training samples in order to counterbalance the preponderance of common features. This algorithm is based on a deep contextualized word representation method named: Embeddings from language models (ELMo), renowned for its power in detecting the finest syntactic and semantic characteristics of words. In a second step, the weight of specific terms is determined by combining two measures: The chi-square weight calculated from the information class provided by labeled data and the semantic weight that corresponds to a score assigned considering the term's association with a severe respiratory illness.

We evaluate the proposed feature selection model using SVM Classifier and the clinical dataset i2b2 [16] enriched with clinical reports gathered from the EMR of the Ibn Sina hospital in Rabat. Experimental results show significant improvement compared to ontology-based feature methods and static word embeddings models with a notable decrease in misclassification rate of test clinical notes related to severe flu by reaching an F-measure of 86.54%.

The principal contributions of this work are: firstly, a novel approach to extend rare and high discriminative words using a contextual word embeddings model. Secondly, a new weighting scheme combining a statistical and a semantic score.

In the remainder of this paper, an overview of related work is provided, followed by a description of our feature engineering approach. Then experimental results are presented and discussed. In the last section, we conclude our work.

II. Related Work

Traditional health surveillance systems rely on epidemiological data collected periodically from various public health system bodies to detect the appearance of a disease [17]. With the emergence of social media and the progressive use of EMRs in healthcare facilities, much health-related data is becoming available to feed automated health surveillance systems with relevant data. In the literature, different approaches have been proposed to take advantage of the textual information available in health-related documents to develop efficient disease prediction systems.

Concerning statistical approaches, SVM, n-gram features, and negation algorithm (NegEx) are experimented in [18] to predict diagnoses from intensive care unit notes. They found that bigrams perform better than other n-gram representations. The negation algorithm does not improve the performance of unigrams. In the study described in [19], the death certificates are classified by type of cancer-causing death. The n-grams and features extracted from the SNOMED CT ontology are employed together to train an SVM classifier, except for certain rare and ambiguous cancers, the proposed model remains effective. In the two previous studies, the authors reported that their classification models are resource-intensive and time-consuming. To overcome this defect, feature selection methods are adopted because they help select relevant features while reducing feature space dimension.

A feature selection approach based on chi-square and t statistical tests is proposed in [20]. It consists of selecting a ranked subset of features from different intensive care unit reports. A configurable threshold determines features list size. A binary classification model is then trained on n-grams, UMLS concepts, and assertion values associated with pneumonia expressions as features to identify pneumonia cases. Experiments show that the number of selected features has no significant impact due to noisy features. In terms of performance t-test, the union of t and chi-square tests and the combination of all feature types provide the best results.

Motivated by the breakthrough of ontologies in the medical domain, many researchers are working to exploit the knowledge presented in ontologies to make predictions on events related to the medical domain. For example, the extended syndromic surveillance ontology is developed in [21]. Its role is to facilitate early disease prediction. It is designed to identify clinical text concepts and then associate the extracted concepts with a particular syndrome. This ontology is created around concepts and their relations, which is tedious and requires excellent domain expertise. Moreover, automated ontologies are conceived in [22] when the proposed model inferred new relations between medical concepts discovered in clinical texts. For this, the model finds its strength in using linked biomedical ontologies to extract relations from enriched concepts.

Despite their power, ontologies are very expensive to setup, because they require domain expertise and are based on standard terminology that changes very little. Therefore they do not take advantage of the explosion of knowledge-rich textual content encapsulated in linguistic forms. An exciting alternative is to use deep neural networks to learn word

embeddings to generate a semantic representation of words in a vector space. In this way, the semantic similarity between words will be determined only by a simple vector computation. Feature selection approaches using neural networks are considered to be useful in selecting the more relevant features. To illustrate this point, a hybrid feature selection method is described in [10] to infer the population's influenza rate. For this, the terms strongly correlated with the target concept are selected from the labeled data. Then, the word2vec language model generates word embeddings for the selected features and retains those with high similarity with the target concept. The problem of rare and out of vocabulary words is addressed in [23], a biomedical word embedding is created by exploiting the subword information. Word embeddings are good at enriching the terminology of existing concepts. They are used in [24] to extend the terminology of dietary supplements. The experimental results prove that the expanded terms are more relevant as search keywords in clinical notes than in external knowledge sources.

Term weighting is an essential step in the classification process. It aims to emphasize useful terms that contribute to better classification accuracy. Traditional methods such as TF-IDF have long proven their effectiveness. Thus, the work presented in [25] elucidates that the use of word2vec word embeddings with TF-IDF is effective for disease classification. In more recent studies, a semantic weight is suggested to express domain relatedness between concepts in the medical domain. We can cite as an example the research work discussed in [26], where word embeddings of all medical concepts are extracted from a corpus of biomedical texts. Then, an association score between each pair of concepts is calculated so that the weight of a concept in a document corresponds to the addition of its TF-IDF frequency with the sum of the association scores of its co-occurring concepts highly associated with it. The proposed weighting scheme outperforms the baseline TF-IDF.

In the literature, several feature-engineering techniques have been proposed. However, to the best of our knowledge, no existing searches explicitly address an approach that emphasizes rare discriminative words in the medical domain. Our work's novelty lies in using a contextual word embeddings model to extend rare features and a new weighting scheme combining statistical and semantic measures.

III. Our Feature Engineering Approach

To alleviate the problem of misclassification when the target classes share several common features, we present a feature enrichment method based on deep neural networks in conjunction with a term weighting scheme in order to strengthen the discriminative power of specific features contained in free-text clinical data. Our approach includes the following steps: Text preprocessing, specific features extraction, word embeddings generation and features weighting. The proposed model is depicted in Fig. 1.

A. Text Preprocessing

Clinical text is full of unnecessary and misspelled words that provide no added value, so before considering feature selection, the text was cleaned up by performing the following actions:

- Text tokenization: Consists of splitting the text into words.
- Text normalization: Consists of representing a word in its canonical form, for example, the words "went" and "going" will be normalized into the word "go".
- Stopwords removal: Words such as prepositions and articles are very common in documents but do not bring useful information for classification.
- Correcting misspelled words: As the dataset contains documents written in French, they are submitted to a spell checker before being translated into English.

B. Medical Concept Extraction

After eliminating stop words, documents are represented by terms that didn't have the same degree of relevance to discriminate between classes. Medical terms are more informative than non-medical terms, but a medical term can be expressed differently depending on the terminology practised by physicians. For example, the rise in body temperature can be designated by one of the terms (Fever, high temperature, hyperthermia). Therefore, a useful term is penalised by the problem of sparsity which reduces its discriminative power. To remedy this problem, the extraction of medical concepts plays a very important role in the normalisation of medical terms into a single synonym denser in documents.

In the context of this work, the MetaMap [27] program developed by the U.S. National Library of Medicine is used to extract medical concepts, its role is to parse the content of a biomedical text in order to recognize medical terms that refer to UMLS concepts.

UMLS organizes the concepts by semantic type, this structure of concepts provided by UMLS is helpful to exclude useless semantic types for prediction, thereby, in consultation with two physicians, we opted for the following semantic types deemed meaningful to predict diseases: Functional Concept, Finding, Virus, Sign or Symptom, Disease or Syndrome, Organic Chemical, Pharmacologic Substance, Medical Device. In Table I, an example of concepts extracted from a clinical text with their semantic types.

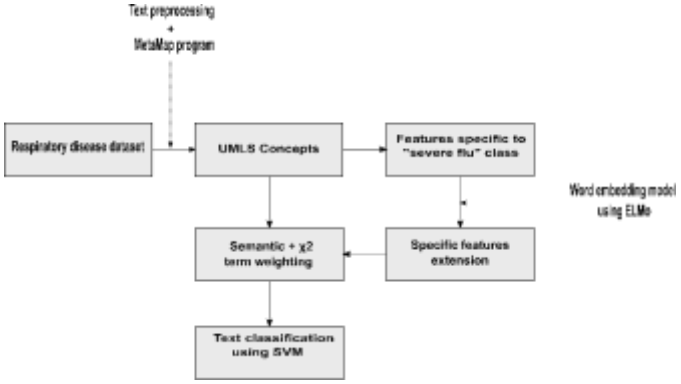


Fig. 1. Architecture of our Feature Engineering Approach.

- **Clinical text:** *"This is a 68-year-old patient with a history of **Type 2 Diabetes** under **ADO** and **chronic smoking** estimated at 30 PA, Admitted for the management of her **respiratory distress** attributed to **pneumonia**".*

TABLE I. EXAMPLES OF UMLS CONCEPTS EXTRACTED FROM A CLINICAL TEXT

Extracted concept	Semantic type
Type 2 Diabetes, Pneuminia	Disease or Syndrome
ADO	Pharmacologic Substance
Chronic smoking	Finding
respiratory distress	Sign or Symptom

C. Word Embeddings Generation

Concept extraction only partially solves the problem of sparsity, medical concepts are also written through abbreviations, acronyms and coding conventions of medical terms specific to each health system that cannot be mapped to UMLS. Furthermore, traditional classifiers cannot retrieve semantic similarities of rare words as they occur in few documents, and therefore have a minor contribution to classification accuracy. In this work, textual documents are to be classified into similar categories, in this case, specific terms are clearly in a minority compared to common terms, yet they are very decisive to predict severe flu. Thus, in order to emphasize these specific terms, documents will be extended with terms similar to them.

Word embedding is a recent technique powered by continuous advancements in deep learning, it is used to learn word vector representation to capture semantic properties helpful to quantitatively estimate the similarity between words.

To optimize the semantic representation of words, word embeddings are often pre-trained on large datasets so that words that occur in the same context have similar meanings.

Word2vec [28] and Glove [29] are among the best known methods, they are called static methods because they only produce a single representation of a word regardless of the context in which the word appears. For example, static methods generate the same representation of the word fever even if its meaning differs depending on the context.

- American election **fever** is approaching.
- **Fever** is a sign of Covid-19.

In this work, Elmo [30] is used to extend the terminology of medial concepts by capturing only the medical meaning of polysemous words. ELMo is a product of Allen NLP, its operating principle is based on two tasks: First, a deep bidirectional LSTM-based language model is pre-trained on a large textual dataset, then, in a second step, the hidden internal states of the model are used to generate the vector representation of words taking into account the context in which the word appears. ELMo can capture the finest syntactic and semantic aspects, and thus outperforms classic models like word2vec et GloVe in many NLP tasks.

In practice, a pre-trained ELMo model trained on PubMed [31] is used to generate 1024-dimensional embeddings of specific terms collected from the training data annotated as "severe flu". For this, it is necessary to identify specific terms that will be submitted to the model, thus, we have defined a specific term as a term whose frequency in the training documents labeled as "severe flu" is clearly higher than its frequency in the training documents labeled as "mild flu", this definition can be formulated as follows:

$$specificTerms = \left\{ t \in Tsf, \frac{D(t,Cmf)}{D(t,Csf)} \leq \alpha \right\} \quad (1)$$

Where:

Csf is the class label of training documents related to "severe flu".

Cmf is the class label of training documents related to "mild flu".

Tsf is the training set of class "severe flu".

D(t, Csf) is the number of documents of class Csf containing the term t.

D(t, Cmf) is the number of documents of class Cmf containing the term t.

$\boldsymbol{\alpha}$ is the threshold which determines the list of specific terms to be selected.

Since ELMo may generate multiple embeddings per word, we average the vectors of all the occurrences of each term to obtain the corresponding word vector.

The i2b2 dataset word vectors are associated with specific term embeddings to form a base of eligible words to extend terminology for specific words. A cosine similarity-based measure is calculated between each specific word and all eligible words, so that except words that reach a similarity above a certain threshold will be retained. In Table II, we list some specific words with their closest similar concepts with a threshold equal to 0.75.

TABLE II. SOME UMLS CONCEPTS WITH THEIR CORRESPONDING MOST SIMILAR WORDS

UMLS concept	Most similar words
Pneumonia	Dyspnea, Desaturation, cyanosis, tachypnea
Swine	H1N1, virus, flu, SRAS, coronavirus, pandemic
distress	ARDS, respiratory, breath, dyspnea
Intubation	Ventilation, nebulization, respirator, ICU
diabetes	Sugar, insulin, hyperglycemia

Although the word *"Swine"* refers to the animal domain, its embedding generated by the model is closely related to the medical context, the same for the word *"Distress"* which is encountered in several contexts but attributed to the medical domain.

D. Term Weighting Scheme

The proposed weighting scheme attempts to assign an appropriate weight to each extracted term and the list of extended features based on their power to discriminate between severe and mild flu. The frequency-based weighting scheme is

not convenient due to the fact that we have at our disposal a labeled training set that tells us about the degree of correlation between terms and the target classes, in addition, even if two terms have a strong correlation with the class "Severe flu", it may not have the same semantic importance for the target category, for example, the term "Pneumonia" is more indicative of a severe case of flu than the term "Fever". Thus, it makes more sense for the proposed weighting scheme to take into account both the degree of correlation with the target class and the semantic importance of terms.

The chi-square test is performed to test the hypothesis of independence between two categorical variables. In text classification, this test is used to rank words from a corpus of textual documents in order to select those that strongly depend on the target class. This dependence between variables is measured by the chi-square test by applying the formula below.

$$\chi^2(d,f,c) = \sum_{i=o}^{1}\sum_{j=0}^{1}\frac{(O_{ij}-E_{ij})}{E_{ij}} \quad (2)$$

Where

O and E are respectively, the observed and expected numbers.

The index i indicates whether the term f is present or not in document d.

The index j indicates whether the document d belongs to class c.

A greater value of the chi-square test indicates that there is a strong correlation between the term and the corresponding class, for that, we retain the chi-square value to calculate the weight of words.

In addition to the weight generated by the statistical correlation between words and the target class, a weight reflecting the semantic importance is used to determine the final weight of a word so as to assign a high semantic weight to terms that are more indicative of severe flu, to do this, the semantic weight of a term corresponds to the cosine similarity between the term and the class "Severe flu", and in order to simplify the calculation of similarity, the class "Severe flu" is represented by the word *"Pneumonia"* since it is often associated with severe complications of flu. In short, the semantic weight is formulated as follows:

$$semantic_w(t) = cosSimilar(emb(t), emb(pneumonia)) \quad (3)$$

Finally, the final weight of the term t is calculated by associating the chi-square weight with the semantic weight according to the formula below:

$$finalWeight(t) = \beta.semantic_w(t) + (1-\beta).\chi^2_w(t) \quad (4)$$

Where

$semantic_w(t)$ is the semantic weight of the term t.

$\chi^2_w(t)$ is the chi-square weight of the term t.

β is a parameter which determines the share of semantic weight in the final weight.

emb(t) is the vector word representation of the term t.

IV. Results and Discussion

A. Datasets

To conduct our experiment, two sources were used to collect clinical documents pretaining to flu.

1) For severe flu cases: By searching with the keyword "Pneumonia", several clinical notes were extracted from the i2b2 dataset, then with the support of two physicians, 500 documents are labelled as "severe flu".

2) For mild flu cases: We collected and translated into english reports of all medical consultations carried out by the pneumology department of Ibn Sina hospital in rabat during the period between January 2017 and February 2020 provided that these consultations did not result in hospitalization. Among the collected documents, 500 reports were annotated as "mild flu".

In the remainder of this section, "Severe flu" documents are considered as positive samples, while "Mild flu" documents are regarded as negative.

B. Evaluation

A health surveillance system must be efficient enough to accurately detect the onset and progression over time of a disease, so our proposed model is designed to meet the following two requirements:

1) Reduce the proportion of mild flu-related documents classified as severe flu, this has the effect of avoiding false outbreak alerts. To assess this performance, the precision which represents the percentage of correct positive predictions over positive predictions is measured as follows:

$$Precision = \frac{T_P}{T_P+F_P} \quad (5)$$

2) Reduce the proportion of severe flu-related documents classified as mild flu in order to prevent an outbreak going undetected. This performance is evaluated through the recall measure which is defined as the percentage of correct positive predictions over the total number of actual positive documents, and it is calculated as below:

$$Recall = \frac{T_P}{T_P+F_N} \quad (6)$$

The proposed classification model is considered sufficiently perfect once high values of precision and recall are reached. In our model, finding a good compromise between precision and recall amounts to determining the values of these two measures which maximize their harmonic mean, also called F-measure. The F-measure is calculated as follows:

$$F-measure = 2.\frac{Precision.Recall}{Precision+Recall} \quad (7)$$

Where

T_P is the number of true positive predictions.

F_P is the number of false positive predictions.

F_N is the number of false negative predictions.

C. Experimental Results and Discussion

Our feature engineering approach is confronted with three baseline methods:

1) Bag of words + TF-IDF: Words resulting from text preprocessing tasks are extracted, then weighted using TF-IDF.

2) UMLS concepts + TF-IDF: Clinical text is mapped to the UMLS concepts using the MetMap program, then the TF-IDF weight is calculated for each extracted concept.

3) The word embeddings model word2vec pretrained on PubMed.

The SVM method is applied to all types of features used in our experience with 90% of the dataset goes to the training set, and the remaining 10% to the testing set. SVM is an ideal choice, as it performs well on different domains [32].

The performance measures of our model depend on two parameters α and β explained in detail in the previous section.

1) α is the threshold below which a term is considered specific to the positive class (Severe flu).

2) β is the share of the semantic weight in the final weight of a term.

When the value of α is very small. The model extracts terms specific to the positive class which rarely occur in the negative class. On the other hand a large value of β means that the semantic weight of a term is greater than its statistical weight.

According to experimental results presented in Fig. 2 the recall value is at its lowest when the model allows selecting as specific term those whose frequency in the positive class is close to their frequency in the negative class, i.e. α close to 1, which can be explained by the fact that the model tends to extend the terminology of common terms that are already dominant over specific terms, causing an increase in FN.

When the value of α decreases, we notice a clear improvement in recall, because the model rejects more common terms, and only those specific to the positive class are extended contributing to rebalance positive training samples by reducing the dominance of common terms. It is to be particularly noted that the recall values peak for α in the range of 0.2-0.3, indicating that most severe flu specific terms are selected when the α parameter value is between 0.2 and 0.3. It is also observed that recall value decrease slightly when α drops below 0.2, which is quite expected since specific terms are very rare and it is unlikely to find terms whose frequency in the positive class is at least 5 times higher than their frequency in the negative class. Except for α equal to 0, the system selects very discriminative terms which exist only in the positive class.

As shown in Fig. 3, the precision varies between 55% and a maximum value of 94.25%. It is minimal when α is equal to 1, then evolves by reaching high values when α is in the interval (0.2-0.3) where the majority of specific words are selected. The precision is relatively high because the number of FP is very low which indicates that the system classifies into the positive category only test documents containing specific terms with very high discriminative power.

The weight of the terms has a significant impact on the model performance. The maximum values of recall and precision are reached when the share of the semantic weight in the final weight is around 60%.When the final weight only includes the chi-square weight, i.e. β= 0, the recall and the precision are respectively 72.45% and 83.48%. On the other hand, a final weight comprising only the semantic weight (β= 1), the recall and the precision take the values 78,35% and 88.62% respectively, which means that the semantic weight contributes more to the discriminative power of the terms.

Although SVM performs well on the most commonly used datasets with an F-measure that typically exceeds 93% [33], SVM classification performance is significantly reduced when trained on our dataset using only the BOW feature representation and the TF-IDF weighting scheme, with an F-measure that barely reaches 53%. Which indicates that traditional feature engineering methods are not effective to discriminate between classes sharing many common features.

The results presented in Table III show that our health surveillance system based on a text classification model constructed through an extension of specific features using ELMo and a weighting scheme combining the semantic and chi-square weights is much better than baseline methods.

Experimental results also show that neural word embeddings models (ELMo and word2vec) are more effective than the bag of words and ontology-based approaches.

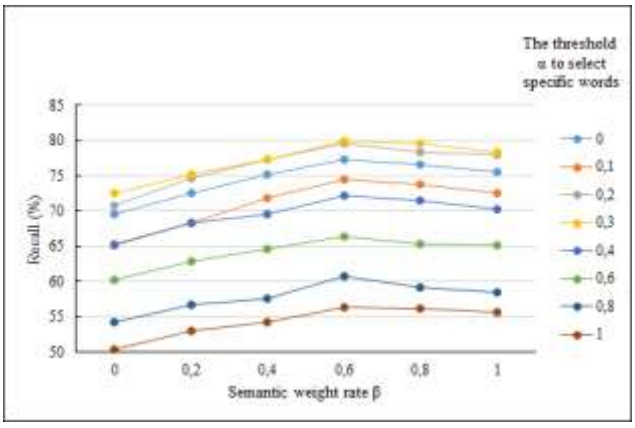


Fig. 2. SVM Classification Recall when Varying Thresholds α and β.

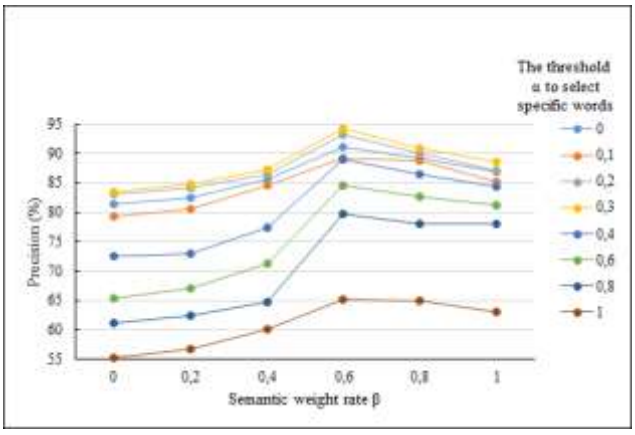


Fig. 3. SVM Classification Precision when Varying Thresholds α and β.

TABLE III. SVM CLASSIFICATION PERFORMANCE OF OUR FEATURE ENGINEERING MODEL COMPARED TO BASELINE METHODS

Features	Recall	Precision	F-measure
Our feature enginnering model	80.00%	94.25%	86.54%
BOW + TF-IDF	52.32%	55.17%	53.70%
UMLS concepts + TF-IDF	65,50%	70,63%	67,96%
Pretrained word2vec	75,23%	80,75%	77.89%

V. CONCLUSION

A system for detecting the occurrence of severe forms of flu by using only clinical texts recorded in EHRs is devised through a text classification model with the challenge of discriminating between severe and mild flu-related documents containing many common features.

To improve classification performance, we have adopted a two-phase approach. In the first phase, with the aim of emphasizing severe flu specific terms deemed rare and discriminative, we have extended these terms by using the pre-trained word embedding ELMo. In the second phase, a combination of two weights is assigned to each term, a semantic weight representing the term's similarity to the word "Pneumonia", and a chi-square weight measuring the correlation between the term and the class "severe flu".

We have found through our experiments that the proposed feature engineering model based on terms extension using deep word representation combined with a weighting scheme that emphasizes discriminative words vigorously improves classification performance when the target classes are very similar.

In this paper, only medical terms are used to determine cases of severe flu. However, opinion words are also useful in deciding the severity of an illness. Thus, mining opinion words occurring in clinical texts is an interesting line of research for our next work.

REFERENCES

[1] P. B. Jensen, L. J. Jensen, et S. Brunak, « Mining electronic health records: towards better research applications and clinical care », Nat. Rev. Genet., vol. 13, no 6, Art. no 6, juin 2012.

[2] S. R. Raman et al., « Leveraging electronic health records for clinical research », Am. Heart J., vol. 202, p. 13-19, 2018.

[3] P. Raghavan, J. L. Chen, E. Fosler-Lussier, et A. M. Lai, « How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? », AMIA Summits Transl. Sci. Proc., vol. 2014, p. 218, 2014.

[4] B. Agarwal et N. Mittal, « Text classification using machine learning methods-a survey », in Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, 2014, p. 701-709.

[5] L. Yu et H. Liu, « Feature selection for high-dimensional data: A fast correlation-based filter solution », in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, p. 856-863.

[6] D. Wang, H. Zhang, R. Liu, et W. Lv, « Feature selection based on term frequency and T-test for text categorization », in Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, p. 1482-1486.

[7] Y. Xu, B. Wang, J. Li, et H. Jing, « An extended document frequency metric for feature selection in text categorization », in Asia Information Retrieval Symposium, 2008, p. 71-82.

[8] P. Samant et R. Agarwal, « Machine learning techniques for medical diagnosis of diabetes using iris images », Comput. Methods Programs Biomed., vol. 157, p. 121-128, 2018.

[9] A. Janecek, W. Gansterer, M. Demel, et G. Ecker, « On the relationship between feature selection and classification accuracy », in New challenges for feature selection in data mining and knowledge discovery, 2008, p. 90-105.

[10] V. Lampos, B. Zou, et I. J. Cox, « Enhancing feature selection using word embeddings: The case of flu surveillance », in Proceedings of the 26th International Conference on World Wide Web, 2017, p. 695-704.

[11] V. N. Garla et C. Brandt, « Ontology-guided feature engineering for clinical text classification », J. Biomed. Inform., vol. 45, no 5, p. 992-998, 2012.

[12] K. Buchan, M. Filannino, et Ö. Uzuner, « Automatic prediction of coronary artery disease from clinical narratives », J. Biomed. Inform., vol. 72, p. 23-32, 2017.

[13] X. Yan et J. Bien, « Rare feature selection in high dimensions », J. Am. Stat. Assoc., p. 1-14, 2020.

[14] K. Patel, D. Patel, M. Golakiya, P. Bhattacharyya, et N. Birari, « Adapting pre-trained word embeddings for use in medical coding », in BioNLP 2017, 2017, p. 302-306.

[15] A. Miaschi et F. Dell'Orletta, « Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation », in Proceedings of the 5th Workshop on Representation Learning for NLP, Online, juill. 2020, p. 110-119.

[16] Ö. Uzuner, B. R. South, S. Shen, et S. L. DuVall, « 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text », J. Am. Med. Inform. Assoc., vol. 18, no 5, p. 552-556, 2011.

[17] G. Shmueli et H. Burkom, « Statistical challenges facing early outbreak detection in biosurveillance », Technometrics, vol. 52, no 1, p. 39-51, 2010.

[18] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, et R. A. Dudley, « N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit », J. Am. Med. Inform. Assoc., vol. 21, no 5, p. 871-875, 2014.

[19] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, et N. Grayson, « Automatic ICD-10 classification of cancers from free-text death certificates », Int. J. Med. Inf., vol. 84, no 11, p. 956-965, 2015.

[20] C. A. Bejan, F. Xia, L. Vanderwende, M. M. Wurfel, et M. Yetisgen-Yildiz, « Pneumonia identification using statistical feature selection », J. Am. Med. Inform. Assoc., vol. 19, no 5, p. 817-823, 2012.

[21] M. Conway, J. Dowling, et W. Chapman, « Developing an application ontology for mining free text clinical reports: the Extended Syndromic Surveillance Ontology », in Proceedings of the Third International Workshop on Health Document Text Mining and Information Analysis, Slovenia (LOUHI 2011), 2011, p. 75-82.

[22] M. Alobaidi, K. M. Malik, et M. Hussain, « Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain », Comput. Methods Programs Biomed., vol. 165, p. 117-128, 2018.

[23] Y. Zhang, Q. Chen, Z. Yang, H. Lin, et Z. Lu, « BioWordVec, improving biomedical word embeddings with subword information and MeSH », Sci. Data, vol. 6, no 1, p. 1-9, 2019.

[24] Y. Fan, S. Pakhomov, R. McEwan, W. Zhao, E. Lindemann, et R. Zhang, « Using word embeddings to expand terminology of dietary supplements on clinical notes », JAMIA Open, vol. 2, no 2, p. 246-253, 2019.

[25] W. Zhu, W. Zhang, G.-Z. Li, C. He, et L. Zhang, « A study of damp-heat syndrome classification using Word2vec and TF-IDF », in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016, p. 1415-1420.

[26] X. Luo et S. Shah, « Concept embedding-based weighting scheme for biomedical text clustering and visualization », in Applied Informatics, 2018, vol. 5, no 1, p. 1-19.

[27] A. R. Aronson, « Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. », in Proceedings of the AMIA Symposium, 2001, p. 17.

[28] T. Mikolov, K. Chen, G. Corrado, et J. Dean, « Efficient estimation of word representations in vector space », ArXiv Prepr. ArXiv13013781, 2013.

[29] J. Pennington, R. Socher, et C. D. Manning, « Glove: Global vectors for word representation », in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, p. 1532-1543.

[30] M. E. Peters et al., « Deep contextualized word representations », ArXiv Prepr. ArXiv180205365, 2018.

[31] Q. Jin, B. Dhingra, W. W. Cohen, et X. Lu, « Probing Biomedical Embeddings from Language Models », ArXiv190402181 Cs, avr. 2019, Consulté le: janv. 28, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1904.02181.

[32] R. Karsi, M. Zaim, et J. El Alami, « Impact of corpus domain for sentiment classification: An evaluation study using supervised machine learning techniques », in Journal of Physics: Conference Series, 2017, vol. 870, no 1, p. 012005.

[33] Z. Liu, X. Lv, K. Liu, et S. Shi, « Study on SVM compared with the other text classification methods », in 2010 Second international workshop on education technology and computer science, 2010, vol. 1, p. 219-222.

An Enhanced Artificial Bee Colony: Naïve Bayes Technique for Optimizing Software Testing

Palak[1*], Preeti Gulia[2], Nasib Singh Gill[3]
Department of Computer Science and Applications
Maharshi Dayanand University, Rohtak, India

***Abstract*—Software driven technology has become a part of life and the quality of software largely depends on the extent of effective testing performed during various phases of development. A wide range of nature inspired searching techniques are employed over years to automate the testing process and provide promising solutions to elude the infeasibility of exhaustive testing. These techniques use metaheuristics and work by converting the problem space into search space. A subset of optimized solutions is searched that reduces overall time by shortening the testing time. Objective: An enhanced Artificial Bee Colony- Naïve Bayes optimizer for test case selection is proposed in this paper. This article also aims to provide brief insights into the emergence of hybrid swarm-inspired techniques over the last two decades. Method: The modified Artificial Bee colony is applied after component selection and further optimization is achieved using Naïve Bayes classifier. The proposed technique is implemented and evaluated taking three benchmark programs into consideration. The proposed technique is also compared to other competitive swarm intelligence-based techniques of its class. Results: The experimental results show that the proposed technique outperforms other swarm-inspired techniques in terms of execution time in a given scenario and capable of higher detection of faults with minimal test case selection. Conclusion: The proposed approach is an improvement over existing techniques and helps in huge time and cost saving. It will contribute to the testing society and enhance the overall quality of the software**.

Keywords*—*Software testing; artificial bee colony; swarm intelligence; Naïve Bayes; test case selection

I. INTRODUCTION

Increasing demand in robust software can be seen as a consequence of rapidly developing hardware industry and outburst in evolution of technology. Smart devices have become our part of our surroundings. Software testing is a very important phase of development of robust software which involves finding possible faults and errors so that the final product meets the overall expectation of the customer without failing. Software testing has always been a hot topic of research for software industry practitioners and researchers. It is the procedure of the identification and authentication of the software services by selecting if it is fulfilling the user's needs. Despite checking correctness of input and output during testing there are many other concerns to deal with. Some of them include dead code, redundant code, faulty components, exceptions etc. These aspects principally affect the overall performance and customer satisfaction. Though, specifications are frequently ignored and there are numerous obstructions to the execution, including the inadequate design, phase limits, absence of automatic apparatuses, and so forth. The cost of testing increases with complexity of the system. More important aspects to be considered are the continuous updates in the system and addition of new functionalities. The configurable architecture of the software now- a- days makes the testing process more difficult due to the behavioral changes of components at each configuration (Myra B. Cohen et al. 2007). Poor testing ultimately leads to system failure and lowers down the faith of the customer. The significance of software testing is that it helps the software programmers for building error-free and acceptable software. A worthy and quality test suite can catch most of the errors without jumping time constraints. The process of testing is carried out at functional (Black Box) as well as structural level (White Box). Functional testing aims to check correctness of the input and output only whereas structural testing checks the deep insights at the root level considering the architectural aspects of the program.

Optimization of testing processes can be done at several levels of software testing life cycle. The field of automated testing is growing day by day due to various underlying benefits. With the increase in dependence of software and ever-growing system requirements, manual testing is not possible at every level of development. Manual testing is time consuming whereas automated testing is fast and repeatable. Smaller projects are feasible to be tested manually but this is not the case for large dynamic projects. The manual testing time rises exponentially with increase in the length of code but this is not the case in automated testing. All you need to do is to write an isolated code called a "Test Code" to test the main functional code. This code can be executed any number of times to find errors whenever required. One more important benefit to mention here is that automated testing permits you to refactor ("changing the structure without altering the actual functionality") the code without any hustle to manually test the code every time you refactor the code. From automated generation of effective test suites to test case selection and prioritization, research is going on which encompasses meta-heuristics and artificial intelligence. Automated testing helps you emphasize more on the quality of the software rather than memorizing what and how to test. Test automation can be done at various levels. Unit testing automation involves testing each independent functional unit without considering the external dependencies. Optimization of such activities utilizes various engineering domains like data mining, artificial intelligence, machine learning, swarm intelligence and many more. Over years, soft computing has emerged as a

*Corresponding Author

promising solution towards optimization problems that involves metaheuristics [1]. Various evolutionary algorithms are also preferred over random search techniques for test suite generation that attracts researchers in this field [2]. This paper also aims to utilize a swarm-based approach for optimization in the testing process.

Rest of the paper is organized as: Section II briefs the emergence of various swarm inspired techniques according to the timeline over the past two decades. Section III presents related work present in different literature over recent years in the similar domain of hybrid Artificial Bee Colony (ABC) optimization techniques. The proposed method in the form of a flowchart is presented in Section IV. Results and evaluation of the research are reported in Section V. Finally, the overall conclusion of the paper is given in Section VI along with the future scope of the article.

II. Related Work

Swarm Intelligence (SI) is a popular field of research that is motivated by the natural phenomenon of a population (group) of various living organisms in their natural habitat for search of food, shelter and security. Over the past few decades swarm-based optimization techniques are emerging at a very fast pace due to the inherent flexibility and robustness [3]. SI refers to collective intelligence that has attracted researchers in almost every area of industry. The community behavior of real living organisms dwelling in nature to protect and feed their community is the real inspiration behind SI. The individuals of a swarm interact mutually with each other and also locally with their surroundings in a decentralized way for survival forming a coherent system that can be modelled into a functional pattern [4]. SI laid its roots in the early 1990s and has become an ever-evolving field since then. This section gives brief glimpses of some important swarm-based optimization techniques that have emerged till date and applied in the field of software testing.

In 1999, “Ant colony optimization (ACO)” was proposed which is inspired by food searching behavior of colonies of real ants [5]. Ants communicate with each other by secreting a chemical substance on their path. The concentration of this chemical increases on the shorter path when the number of ants taking that path increases over time. This simple but robust behavior gave the inspiration to build a meta-heuristics model that can be used in various search optimization problems. ACO is well utilized in test case generation, selection and prioritization problems in past few years [6] [7] [8]. The problem of testing optimization is first converted to graphical search problem and then ACO is applied [9]. Various hybrid approaches with ACO have also been proposed with other techniques such as Genetic Algorithm (GA) which aims to select a minimal test suite for higher fault coverage [10].

An efficient algorithm inspired from the social behavior of bees in the search of food was given by Dervis Karaboga et al. in [11] named as “Artificial Bee Colony (ABC)” Optimization. They considered three types of bees and converted their behavior to a mathematical model. Initially half of the bees in the beehive are termed as “Employed Bees”. They search for the food randomly near the hive and come back to the hive. They dance in front of the second set of the bees called “Onlooker Bees”. This dance serves as the probability function for comparison and selection of the better food source. If the food source of any Employed Bee is exhausted then it becomes the scout and serves as the stopping criteria for the algorithm.

Due to its lightweight deployment with very small amounts of controller factors, numerous hard works have been done to discover ABC research. ABC has gained popularity since its origin and researchers are more interested in making hybrid algorithms that provide more diversification in searching the solution. ABC is inspired from natural behavior of honeybees in the search of nectar and their community behavior in maintaining the highest nectar collection. The success of ABC can be anticipated by vast literature available under reputed indexing that shows the interest of researchers in this approach. Originally the ABC technique employs three types of bees: Employed, Onlooker and Scout bees [11]. The employed bees are linked to a definite food source. Initially one employed bee is assigned to a food source. They transmit vital information such as navigation information, location and the profitability of the food source and carry the data with the rest of bees at the beehive. The onlooker bees are accountable for food source detection exploiting the information delivered by employed bees. The scout bees dispensed randomly to hunt the new food source whenever there is no further improved solution is found by either employed or onlooker bees [D. Karaboga, 2005]. The assumption is that the employed bees whose food source is exhausted are transformed into “scout bees" and commence a new exploration for the food source. The parallel conduct of these three bees speeds up the generation of feasible independent paths and software test suite optimization. ABC performs competitively to other conventional soft computing techniques and has gained popularity over last decade due to its easy implementation. Various hybrid and enhanced ABC techniques evolved over the past decade that are used for optimization problems especially in the field of software testing. Table I provides a brief insight into such hybrid ABC techniques:

TABLE I. EXISTING HYBRID ARTIFICIAL BEE COLONY BASED OPTIMIZATION TECHNIQUES

Author	Year	Technique used	Application Area
[Lakshminarayana P et al. [12]**]**	2021	Hybrid Cuckoo Search and Bee Colony Algorithm	Optimization of test cases and generation of path convergence within
[Hussain, Kashif et al. [13]**]**	2020	Scoutless ABC	Model-driven testing
[Saju Sankar S et al. [14]**]**	2020	Comprehensive Improved Ant Colony Optimization (ACIACO)	Automated test case generation
[Ammar K. Alazzawi et al. [15] [16] **]**	2019, 2020	Hybrid artificial bee colony algorithm and practical swarm optimization with constraint support	Generation of variable t-way test sets
[Snehlata Sheoran et al. [17]**]**	2019	Memory based ABC	Data flow testing to find out and prioritize the definition-use paths
[Hu Peng et al. [18]**]**	2019	Best Neighbor-guided artificial bee colony	Continuous optimization problems
[Sandeep Dalal et al. [19]**]**	2018	BCO-m-GA	Test case selection
[Faten Hamad [20]**]**	2018	Modified ABC	Software structural testing
[Sahoo, Rajesh et al. [21]**]**	2017	Hybrid PSO and BCA	Model-driven testing
Zohreh Karimi Aghdam et al. [22]	2017	Modified Fitness Function in ABC	Generate Test Data for Software Structural Testing
Xianneng Li et al. [23]	2016	Artificial bee colony algorithm with memory	Continuous optimization problems
D. Karaboga et al. [24]	2014	Quick ABC with different functions for employed and onlooker bees	Numerical Optimization Problems

III. PROPOSED APPROACH (ENHANCED ABC- NAÏVE BAYES OPTIMIZATION)

In this section, a novel "Enhanced ABC- Naïve Bayes Optimization (ABC-NB)" is proposed for software test case selection. Fig. 1 shows the flowchart of proposed methodology that is inspired from memory-based ABC [17], [19], [23] along with the Naïve Bayes Classifier to further enhance the results.

ABC is highly exploited in the field of software testing that shows the capability of the method. ABC also provides the inherent advantage of independent and parallel behavior of three types of honey bees. Also, this is a non- pheromone-based technique that decreases the computational complexity up to a great extent [25]. That's why we prefer ABC over other swarm-based techniques for optimization of the testing process. The algorithm starts with the selection of the project. We are considering Components Based Software (CBS) development paradigm into account due to the inherent modularity and capability of handling complex projects.

The proposed approach works as follows: component-based projects are selected and uploaded to the repository and their individual components are extracted. Here components refer to each individual unit of work that has predefined interfaces and boundaries. Further each component is subdivided into modules. A component may consist of one of more modules and other components. "Enhanced ABC with memorizing capability" is applied for selecting a subset of test cases in the given fault matrix. The memory element is used to store the best solution found so far to maintain overall intensification as well as diversification. Originally ABC has three phases each related to three different types of bees in the beehive. This behavior is inspired from real beehives where nectar collection is a result of highly organized and collaborated team work. Fig. 1 shows the detailed flowchart of the proposed technique. The various phases and role of different type of bees is as follows:

A. Initialization

First of all, we need to initialize the population size i.e., no. of candidate solutions (initial number test cases in our case) that is denoted by TN. Each solution (test case) is related to D dimensional parameter vector that defines a particular solution based on fault matrix i.e.

Xi = $\{x_i^1, x_i^2, \ldots, x_i^D\}$, i = 1, 2, . . ., TN.

Initially the memory element is kept empty.

For a fault j in fault matrix for ith test case, the initial value x_i^j is generated by

$$x_i^j = x_{min}^j + \text{rand}(0, 1) \times (x_{max}^j - x_{min}^j) \quad (1)$$

where, i = 1, 2, . . ., TN and j = 1, 2, . . ., D. rand(0, 1) is a random number whose value belongs to [0, 1], max and min are the maximum and minimum value in case of each parameter respectively.

B. Employed Bees

Each employed bee maintains individual solutions so their number is equal to the total number of test cases, that is, TN. For each test case i, employed bees generate a new vector Y_i.

The neighbor search is performed by modifying jth parameter of Y_i where j ∈ {1, 2, . . ., D} is selected randomly. The following equation is used for updates done by employed bees (Karaboga and Basturk 2007):

$$y_i^j = x_i^j + \Phi_i^j \times (x_i^j - x_k^j). \quad (2)$$

Here k is randomly selected and i ≠ k. X_i will be replaced by Y_i; in the population if Y_i is better. Φ_i^j is for randomness ranging in [−1, 1].

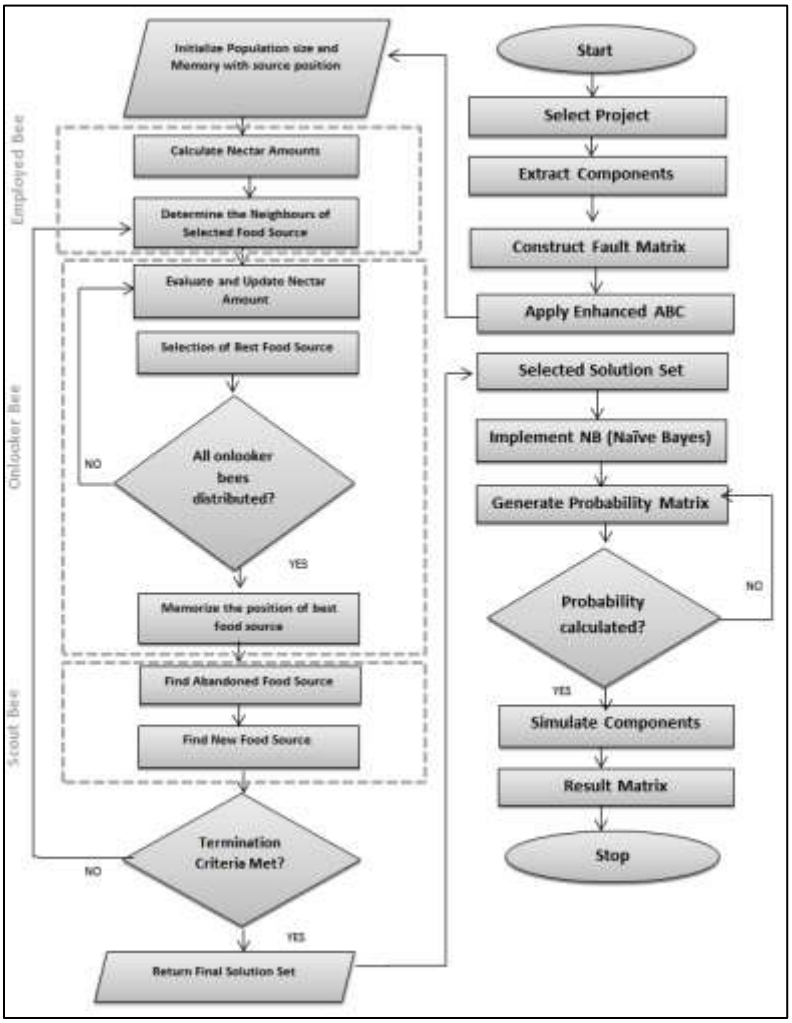


Fig. 1. Flowchart of Proposed Methodology.

Each employed bee also has memory ME_i, that stores the best solution found so far. After every iteration memory is also updated. During next cycle, ME_i is searched first before randomly selecting neighbor solution.

C. Onlooker Bees

The probability of selecting a test case "i" by an onlooker bee is denoted by p_i, which is calculated by

$$p_i = \frac{fit_i}{\sum_{j=1}^{TN} fit_j} \quad (3)$$

where fit_i denotes the fitness value of ith test case, which is calculated on the basis of probability of a test cases to find given set of errors. The onlooker bee also generates a new solution Yi using equation (2) similar to the employed bee. Each onlooker bee also has memory MO_i, that stores the best solution found so far. After every iteration memory is also updated. During next cycle, MO_i is searched first before randomly selecting neighbor solution.

D. Scout Bee

When a solution cannot be further improved by either employed or onlooker bee that solution is considered as poor performing in the process of evolution as must be removed from the final solution set. In such a scenario, a scout bee is generated, it abandons the poor performing test cases and starts with a whole new random solution. Scout bees maintain randomness and diversification in the algorithm.

After the application of Enhanced ABC, a solution set of promising test cases is returned to the system. Here comes the role of Naïve- Bayes Classifier. It generates the probability matrix over the solution set that is returned in the previous stage and further classifies the solution set. Hence a reduced result set is generated. Naïve Bayes is a family of classification techniques that assumes all features into consideration as independent and of equal weight. The proposed technique is applied on three component-based student projects and implemented in ten iterations with fault matrix of size 50*50 in each project. Errors are induced using mutation to test the efficiency of the proposed method.

IV. Results and Discussion

The proposed approach is implemented in "Visual C# Express 2010" using three student's projects namely: Café Management (CM), Hospital Management (HM), and Payroll System (PS). All of them are implemented in C# using component-based paradigm. The reason for selecting Visual

C# projects is the intrinsic component-based approach that is offered by this platform. The details of these projects are given below in Table II:

TABLE II. DETAILS OF STUDENT'S PROJECTS

Project Name	kLOC	Number of Components	Total Number of Modules
Café Management (CM)	69	6	46
Hospital Management (HM)	78	5	53
Payroll System (PS)	43	3	34

A. No. of Selected Test Cases vs No. of Faults Detected

The experiment is conducted for ten iterations to rule out any chances of error and for averaging of the results with fault matrix of size 50*50 in each project. Initially it is assumed that each test case is capable of finding at least one error. Errors are induced using mutation to test the efficiency of the proposed method. Gradually as the algorithm converges, a smaller fault matrix with a lesser number of selected test cases, Table III shows the performance of the proposed technique in terms of percentage of test case selected and percentage of faults detected.

Fig. 2 shows the results in graphical form. It is depicted that the proposed ABC-NB technique selects less than 47 % of test cases to achieve near optimal fault coverage. The size of the test suite the faster the process is. The results of the proposed technique prove promising in selecting better and shorter test suite so that overall execution time can be reduced.

B. Comparison of Execution Time

Being a costly and time-consuming process, software test execution time plays a very important role. Cost can be greatly minimized by decreasing the execution time without compromising with the quality of test suite. On the basis of fault matrix, the execution time of selected test cases by the proposed approach is compared with the execution time of selected test cases by other swarm-based techniques namely PSO, ACO, ABC and the results are summarized in Table IV.

Fig. 3 shows the comparative graph for the same depicting the clear time saving that can be achieved using the Enhanced ABC- Naïve Bayes technique. It can be argued from the experimental results that the proposed hybrid technique is capable of providing time saving as compared to other competitive techniques. As it shortens the execution time in the given scenario, efforts and cost are automatically reduced.

TABLE III. PERFORMANCE OF ENHANCED ABC- NAIVE BAYES FOR SELECTION OF TEST CASES AND FAULTS COVERED

Project name	Total No. of Test Cases	Total Number of faults	Number of test cases selected	% of test cases selected	No. of Faults Detected	% of faults detected
CM	50	50	21	**42%**	49	**98%**
HM	50	50	19	**38%**	50	**100%**
PS	50	50	23	**46%**	50	**100%**

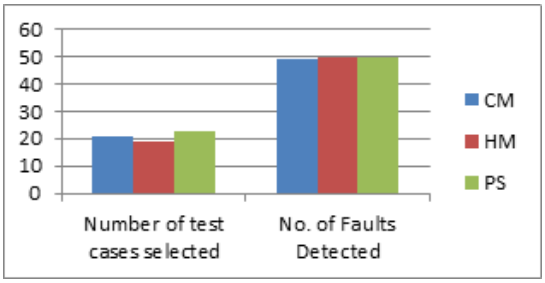


Fig. 2. Performance of Enhanced ABC- Naive Bayes.

TABLE IV. EXECUTION TIME OF SELECTED TEST CASES IN MILLISECOND (MS)

Algorithm/ Project	CM	HM	PS
PSO	165	133	124.3
ACO	160	145.5	120.4
ABC	**125.6**	143	111.9
Enhanced ABC- Naive Bayes	127.3	**129.4**	**105.4**

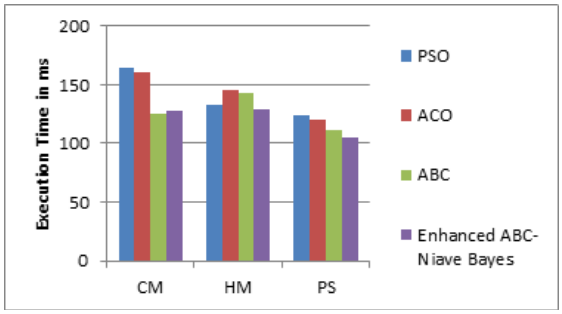


Fig. 3. Comparison of Execution Time.

V. CONCLUSION AND FUTURE SCOPE

Swarm intelligence always inspired researchers to optimize the search problems to save time and money. In this paper, a novel Enhanced ABC – Naïve Bayes algorithm is proposed that is inspired from the colony of honey bees for optimization of test case selection. Being a time consuming and important task, testing always requires optimization. The proposed technique is applied on three component-based student projects and implemented in ten iterations with fault matrix of size 50*50 in each project. Errors are induced using mutation to test the efficiency of the proposed method. The results show that the proposed method is able to find near optimal (i.e.,~ 100%) faults in less than 47 % of total test cases. Thus, a huge amount of time saving can be achieved. The proposed method ABC-NB is also compared with other swarm-based techniques of its class by taking execution time of the selected test cases as a parameter. The proposed technique outperforms PSO, ACO and original ABC as depicted by the results. In future, the proposed method will be compared and evaluated with other swarm-based techniques of its class using more parameters to assess the efficiency and accuracy of the proposed method.

REFERENCES

[1] P. Gulia and P. Palak, "Nature Inspired Soft Computing Based Software Testing Techniques For Reusable Software Components," J. Theor. Appl. Inf. Technol., vol. 95, no. 24, pp. 6996–7004, 2017.

[2] J. Campos, Y. Ge, N. Albunian, G. Fraser, M. Eler, and A. Arcuri, "An empirical evaluation of evolutionary algorithms for unit test suite generation," Inf. Softw. Technol., vol. 104, no. August, pp. 207–235, 2018, doi: 10.1016/j.infsof.2018.08.010.

[3] A. Abraham, H. Guo, and H. Liu, "Swarm Intelligence: Foundations, Perspectives," Swarm Intell. Syst., vol. 25, pp. 3–25, 2006.

[4] I. Aydogdu, M. P. Saka, and E. Do, "Analysis of Swarm Intelligence À Based Algorithms for Constrained Optimization," in Swarm Intelligence and Bio-Inspired Computation, Elsevier Inc., 2013, pp. 25–48.

[5] M. Dorigo and G. Di Caro, "Ant Colony Optimization: A New Meta-Heuristic," in Proceedings of the 1999 congress on evolutionary computation-CEC 99, 1999, pp. 1470–1477.

[6] B. Suri and S. Singhal, "Analyzing test case selection & prioritization using ACO," ACM SIGSOFT Softw. Eng. Notes, vol. 36, no. 6, p. 1, 2011, doi: 10.1145/2047414.2047431.

[7] U. M Diwekar and B. H Gebreslassie, "Efficient Ant Colony Optimization (EACO) Algorithm for Deterministic Optimization," Int. J. Swarm Intell. Evol. Comput., vol. 05, no. 01, 2015, doi: 10.4172/2090-4908.1000131.

[8] P. Palak and P. Gulia, "Ant Colony Optimization Based Test Case Selection for Component Based Software," Int. J. Eng. Technol., vol. 7, no. 4, pp. 2743–2745, 2018, doi: 10.14419/ijet.v7i4.17565.

[9] S. F. Ahmad, D. K. Singh, and P. Suman, "Prioritization for Regression Testing Using Ant Colony Optimization Based on Test Factors," in Intelligent Communication, Control and Devices, 2018, pp. 1353–1360.

[10] P. Palak and P. Gulia, "Hybrid swarm and GA based approach for software test case selection," Int. J. Electr. Comput. Eng., vol. 9, no. 6, pp. 4898–4903, 2019, doi: 10.11591/ijece.v9i6.pp49898-4903.

[11] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," J. Glob. Optim. 39, pp. 459–471, 2007, doi: 10.1007/s10898-007-9149-x.

[12] L. P and T. V Suresh Kumar, "Automatic Generation and Optimization of Test case using Hybrid Cuckoo Search and Bee Colony Algorithm," J. Intell. Syst., vol. 30(1), pp. 59–72, 2021.

[13] K. Hussain, M. Najib, M. Salleh, S. Cheng, Y. Shi, and R. Naseem, "Artificial bee colony algorithm: A component-wise analysis using diversity measurement," J. King Saud Univ. - Comput. Inf. Sci., vol. 32, no. 7, pp. 794–808, 2020, doi: 10.1016/j.jksuci.2018.09.017.

[14] S. S. S. B and V. C. S. S. B, "An Ant Colony Optimization Algorithm Based Automated Generation of Software Test Cases," in the International Conference on Swarm Intelligence. (ICSI 2020), 2020, vol. 1, pp. 231–239, doi: 10.1007/978-3-030-53956-6.

[15] A. K. Alazzawi, H. Rais, and S. Basri, "HABC : Hybrid Artificial Bee Colony For Generating Variable T-Way Test Sets," J. Eng. Sci. Technol., vol. 15, no. 2, pp. 746–767, 2020.

[16] A. K. Alazzawi, H. Rais, S. Basri, and Y. A. Alsariera, "PhABC : A Hybrid Artificial Bee Colony Strategy for Pairwise test suite Generation with Constraints Support," in IEEE Student Conference on Research and Development (SCOReD), 2019, no. October, pp. 106–111, doi: 10.1109/SCORED.2019.8896324.

[17] S. Sheoran, N. Mittal, and A. Gelbukh, "Artificial bee colony algorithm in data flow testing for optimal test suite generation," Int. J. Syst. Assur. Eng. Manag., vol. 11, pp. 340–349, 2019, doi: 10.1007/s13198-019-00862-1.

[18] H. Peng, C. Deng, and Z. Wu, "Best neighbor-guided artificial bee colony algorithm for continuous optimization problems," Soft Comput., vol. 1, no. 23, pp. 8723–8740, 2019, doi: 10.1007/s00500-018-3473-6.

[19] S. Dalal, "Performance Analysis of BCO-m-GA Technique for Test Case Selection," Indian J. Sci. Technol., vol. 11(9), no. March, 2018, doi: 10.17485/ijst/2018/v11i.

[20] F. Hamad, "Using Artificial Bee Colony Algorithm for Test Data Generation and Path Testing Coverage," Mod. Appl. Sci., vol. 12, no. 7, pp. 99–112, 2018, doi: 10.5539/mas.v12n7p99.

[21] R. Sahoo, S. Nanda, and D. P. Mohapatra, "Model Driven Test Case Optimization of UML Combinational Diagrams Using Hybrid Bee Colony Algorithm," Int. J. Intell. Syst. Appl., no. July, 2017, doi: 10.5815/ijisa.2017.06.05.

[22] Z. K. Aghdam and B. Arasteh, "An Efficient Method to Generate Test Data for Software Structural Testing Using Artificial Bee Colony Optimization Algorithm," Int. J. Softw. Eng. Knowl. Eng., vol. 27, no. 6, pp. 951–966, 2017, doi: 10.1142/S0218194017500358.

[23] X. Li and G. Yang, "Artificial bee colony algorithm with memory," Appl. Soft Comput., vol. 41, pp. 362–372, 2016, doi: 10.1016/j.asoc.2015.12.046.

[24] D. Karaboga and B. Gorkemli, "A quick artificial bee colony (qABC) algorithm and its performance on optimization problems," Appl. Soft Comput. J., vol. 23, pp. 227–238, 2014, doi: 10.1016/j.asoc.2014.06.035.

[25] S. Sekhara, B. Lam, M. L. H. Prasad, U. K. M, and S. Ch, "Automated Generation of Independent Paths and Test Suite Optimization Using Artificial Bee Colony," in International Conference on Communication Technology and System Design, 2011, vol. 30, no. 2011, pp. 191–200, doi: 10.1016/j.proeng.2012.01.851.

Intelligent Climate Control System inside a Greenhouse

A. Labidi[1], A. Chouchaine[2], A. Mami[3]
University of Tunis El Manar, Faculty of Science, UR 17ES11 LAPER, 2092 Tunis, Tunisia

***Abstract*—An agricultural greenhouse is an environment to ensure intensive agricultural production. The favorable climatological conditions (temperature, lighting, humidity ...) for agricultural production must be reproduced in a non-natural way by controlling these parameters using several actuators (heating/air conditioning, ventilation, and humidifier/ dehumidifier). The objective of this study is to control the humidity inside the greenhouse; it is a problem that remains to be negotiated. To that end, an actuator based on a humidifier and a dehumidifier was installed in an experimental greenhouse and activated by a fuzzy logic controller to achieve the desired optimal indoor humidity in the greenhouse.**

***Keywords*—Greenhouse; climate; humidity; fuzzy logic controller; humidification; dehumidification**

I. Introduction

A greenhouse is intended to protect plants and promote greenhouse production by creating climatic conditions that are more favorable than the local climate. Therefore, the regulation of the indoor climate, in particular the relative humidity, is necessary to create an ideal environment for plant growth. Greenhouse climate management is controlled by different actuators: ventilation, heating, humidification, and dehumidification system. Many different sensors can measure the actual climate conditions for a precise real-time control method. The need to improve these climatic conditions has called for advanced control algorithms due to the complexity and nonlinearity of the greenhouse system. Many researchers have developed several control strategies to improve the indoor microclimate such as Proportional - Integral - Derivative controller (PID controller) [1], Neural Network [2,3], the PI controller (SSODPI and PI-SSOD event controllers) [4], Adaptive Neuro-fuzzy controller [5,6,7,8], Genetic algorithm [9], Optimal control [10], Predictive Neural Control [11], four control techniques have been developed [12]: Adaptive Neuro-Fuzzy Control (ANFIS), Fuzzy Logic Control (FLC), PI Control and Artificial Neural Network Control (ANN), to adjust the temperature inside the greenhouse, and a Fuzzy Logic Controller (FLC) [13,14] which is a valuable element in the control of hardly identifiable and non-linear systems. Also, several studies have established the importance and usefulness of the FLC controller and its tool to solve the problem of complexity and non-linearity of the greenhouse system [15] from which presented a comparative study of a basic fuzzy controller and optimized fuzzy controllers to show their advantages and disadvantages. The author in [16] have developed a fuzzy modeling application to control the indoor air temperature of a MISO greenhouse, [17] have used this application with a new approach that automatically organizes a fuzzy flat system into a hierarchical collaborative architecture, this architecture adapted to transfer the information contained in the fuzzy rule sets to another fuzzy subsystem.

In this paper, an FLC is developed to manage humidity combined with a humidifier and a dehumidifier to ensure the optimization of the microclimate. The importance of this work comes from the fact that humidity is the most difficult environmental factor to control in an agricultural greenhouse, and whether the humidity level is too high or too low; the loss of quality decreases the selling price of the crops and increases the production costs, thus reducing profits. This work is organized as follows; the second section deals with a description of the experimental set-up studied with the measuring equipment followed by dynamic modeling of the greenhouse's internal moisture behaviors. The third section describes the Fuzzy Logic Controller strategy that was developed and applied to the greenhouse to improve the humidity inside the greenhouse. In the fourth section, a presentation of the results is given followed by a discussion. Finally, this study will be complemented by a general conclusion.

II. Dynamic Model of the Agricultural Greenhouse

A. Greenhouse Modeling

The dynamic modeling of the agricultural greenhouse presents the balance of energy and mass exchanges inside the microclimate. The study of such a model can be useful to observe the actuator's influence on the system's behavior containing two differential equations describing the energy balance of the indoor humidity and the indoor air temperature. According to the law of energy conservation, the equations describing the heat balance of the indoor air can be obtained from the following equation [18].

$$C_z \frac{dT_{in}(t)}{dt} = (Q^{Solar\ radiation} + Q^{ht,cover} + Q^{ht,ground} + Q^{Heaters} + Q^{Humidifying} + Q^{ventilation}) \quad (1)$$

With :

C_z : The thermal capacity of the air (product of air specific thermal capacity, air density and greenhouse volume) and all other elements thermal equilibrium with the air.

- $Q^{Solar\ radiation}$: heat gain by solar radiation (Wm^{-2}).
- $Q^{ht,coverture}$: heat transfer from the envelope between the inside and outside of the greenhouse [19]. (Wm^{-2}).

- $Q^{ht,ground}$: heat transfer to the ground (Wm^{-2}).
- $Q^{Heaters}$: the heat gain of the heating system (Wm^{-2}).
- $Q^{Humidifying}$: the cooling effect of the humidification system(Wm^{-2}).
- $Q^{Ventilation}$: heat loss due to ventilation (Wm^{-2}) [20,21].

Indoor air is also characterized by its relative humidity. The indoor humidity equation can be expressed by:

$$\frac{dHin}{dt} = H^{Evapotranspiration} + H^{Humidifying} - H^{Dehumidifiying} - H^{Ventilation} - H^{Condensation} \quad (2)$$

Avec :

- $H^{Evapotranpiration}$: vapour transferred from the ground to the indoor air by evapotranspiration (gs^{-1} of water).
- $H^{Humidifying}$ and $H^{Déshumidification}$ are, respectively, the humidifying and dehumidifying rates : (gs^{-1} of water).
- $H^{Ventilation}$: water exchanged by ventilation (gs^{-1} of water).
- $H^{Condensation}$: the water condensation process (gs^{-1} of water).

B. Experimental Setup

The agricultural greenhouse used has a transparent plastic cover (PVC), it is located north of Tunis with its axis parallel to the east-west direction. The process is surrounded by a 2.5 m high wind breeze to reduce the influence of the wind. The experimental device is 1.5 m long with a width of 1 m and a height of 1.15 m (Fig. 1) and is equipped with a data acquisition and processing system. A humidifier is placed inside the greenhouse. A dehumidifier type VERODRY 2009 LCD was also used; its role is to decrease the humidity level. A data acquisition system with several probes is used to measure different climatic parameters such as temperature, humidity, and solar radiation inside and outside the greenhouse; two sensors type LM35CZ are used to measure temperature, two sensors type SY-230 to measure relative humidity. The sensor's technical characteristics are listed in Table I [22,23,24]. The electronic components are protected by naturally ventilated boxes inside and outside the greenhouse (Fig. 2).

A thermopile cell type LPYRA03 is used to measure solar radiation. A data acquisition, processing, and control board type STM32F407VG Discovery is used to take measurements of different climatic parameters. The signals delivered by the sensors are amplified and conditioned using instrumentation amplifiers type AD620 before being transferred to the acquisition board (Fig. 3). The "STM Studio" software is used to process and transfer the data from the acquisition board to the computer; these data are measured every 30 seconds day and night then stored in files constituting a fairly large database of the greenhouse in different climatic conditions.

Fig. 1. Experimental Greenhouse.

TABLE I. SENSORS CHARACTERISTICS

	Temperature sensor LM35	**Humidity sensor SY-230**	**Thermopile cell LPYRA**
Temperature range	-55ºC to 150ºC		-40 ºC to 80 ºC
Rated voltage	4 to 30 volt	Dc 5.0 volt nominal voltage	
Rated power	< 60 µA	≤ 3.0 mA Nominal power	
Operating humidity		10 – 90 % RH	
Measuring range			0 to 2000 W/m^2
Typical sensitivity			10 µV/(W/m^2)
Impedance	Low impedance output 0.1 Ω for 1mA load		33 Ω to 45 Ω
Field of view			2π sr
Spectral field			305 nm to 2800 nm

Fig. 2. Humidity Sensor. [25].

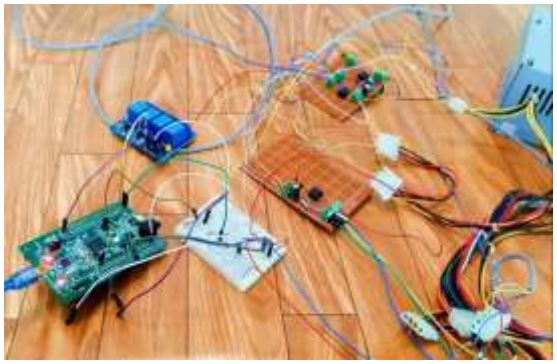


Fig. 3. Data Acquisition and Control Board.

III. Fuzzy Control Scheme

A. Fuzzy Logic Concept

Climate regulation is of vital importance for the plant grown in the greenhouse. In this research work, a methodology based on fuzzy logic control is used. This particular technique is a valuable element in the control of systems whose parameters are subject to significant variations. Fuzzy logic control uses inferences with several rules linking input fuzzy variables to an output fuzzy variable so the fuzzy model is generally composed of fuzzification which involves a transformation of input variables into linguistic. Each input has its group of membership functions. The inference step consists of defining a logical relationship between the system inputs and outputs in the form of membership rules which can be drawn up in an inference table. The de-fuzzification step is the inverse of the fuzzification step, it allows to reconvert the fuzzy output into a net output (I can't find any other synonym for net) (Fig. 4).

B. Humidity Fuzzy Control

We have focused our study on indoor relative humidity control. This study is done because of the importance of humidity in greenhouse agricultural production. Humidity control was ensured by two actuators: the humidification system and the dehumidification system.

The fuzzy logic controller was developed to regulate the humidity inside the studied greenhouse where the included inputs are the error (between the set-point and the indoor humidity) and the previous control action of the actuators.

The linguistic terms used to describe the values of the inputs were negative (neg), null, positive (posi). The linguistic terms for the previous control action are dehumidification (dehumid), no_action and humidification (humid). The linguistic terms for the fuzzy control output are mf1, mf2 and mf3. The outputs are described by three levels where each one is associated to a different value; 1 is associated to the humidifier activation, the dehumidifier activation is associated to -1 and zero when both actuators are not activated. The same values are used in the input of the fuzzy controller since it represents the previous control action, (see Fig. 5, 6 and 7).

Initially, the inputs in each rule are fuzzified; their values are used as inputs of the membership functions belonging to each rule, the result are then used in a product method. The weight of each rule is obtained from each product.

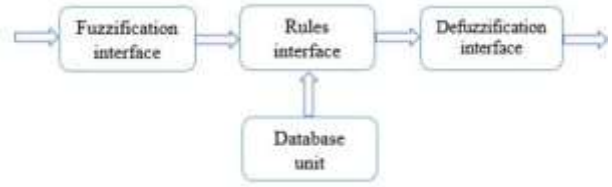


Fig. 4. Block Diagram of the Fuzzy Regulator.

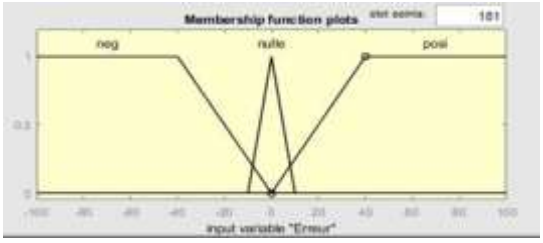


Fig. 5. Membership Functions of the Input Error.

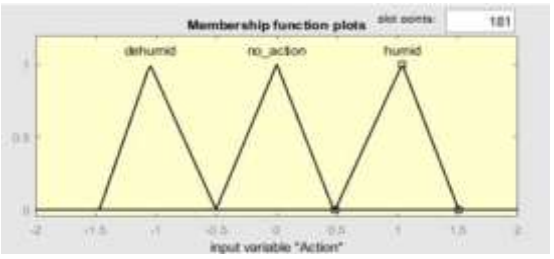


Fig. 6. Membership Functions of the Input Action.

Fig. 7. Membership Functions of the Output.

It is worth noting that the defuzzification process is not applied here. That is deeply related to the On/Off activation way used by the actuators. Instead, the result of the inference process is used to determine the maximum weight between all the rules then apply the associated control output to the actuators (see Table II). The fuzzy controller uses nine rules.

TABLE II. Fuzzy Rules of the Relative Humidity

Error	Action	Command
Humid	Posi	Mf3= 1
Humid	Null	Mf3= 1
Humid	Neg	Mf1= -1
No_action	Posi	Mf3= 1
No_action	Null	Mf2= 0
No_action	Neg	Mf3= 1
Dehumid	Posi	Mf3= 1
Dehumid	Null	Mf1= -1
Dehumid	neg	Mf1= -1

IV. Results and Discussion

The samples used for climate control were taken from 14/10 to 16/10 at the LAPER laboratory. These measurements include solar radiation and humidity inside and outside the greenhouse.

Fig. 8 shows the solar radiation measured for about 46 hours, while the indoor and outdoor humidity is shown in Fig. 9 measured at the same time.

The responses of the two sensors (internal and external humidity) follow a periodic phenomenon of humidity variation and show almost identical evolutions with some differences. The external humidity depends on the solar radiation, for this reason, it increases during the night. Likewise, the indoor humidity increases during this period to approach that of the outside (and exceeds (80%), and this excess requires the use of the dehumidifier to lower its value to a level suitable for the plant. On the contrary case, during the day the interior humidity decreases until it reaches its minimum value (20%) that requires the use of a humidifier to increase the humidity.

The main objective of the control is to gently force the humidity inside the greenhouse to follow their desired trajectory. The response of the greenhouse interior humidity is shown in Fig. 10 for a period of 46 hours from 22/10 to 24/10 where the relative humidity set point was equal to 60%. The activation of the humidification is associated to 1, that of the dehumidification is associated to -1 and 0 describes the deactivation of the two actuators.

It can be seen that the humidity level during the day and night is maintained around its 60% set point except when the application time is between 22 and 24 hours. At that time, the humidification operates continuously but the humidity response cannot keep up. That can be explained by the high level of solar radiation reached in those moments. Apart from that period, one can observe that the dehumidifier system is more active in the night when the indoor humidity tends to surpass the set point. In the daytime, the humidifier system is more active or else the indoor humidity will decrease under the level of the set point.

The relative humidity evolution inside the studied greenhouse without and with the fuzzy controller is shown in Fig. 9 and 10, respectively. The fuzzy controller shows satisfactory results by better stability which proves the efficiency of this controller hence this study aimed to show the performance of FLC for setpoint monitoring and disturbance rejection.

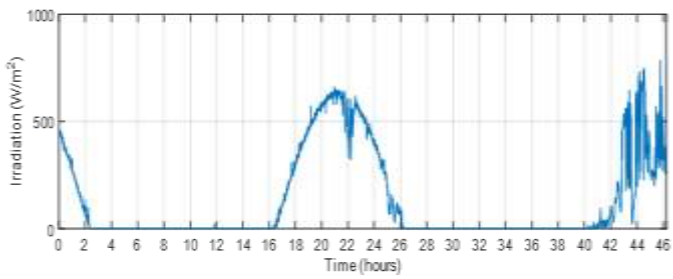


Fig. 8. Variation of the Intercepted Solar Radiation.

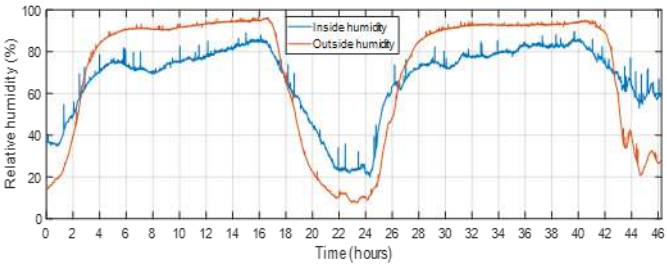


Fig. 9. Indoor and Outdoor Humidity Behavior without Control.

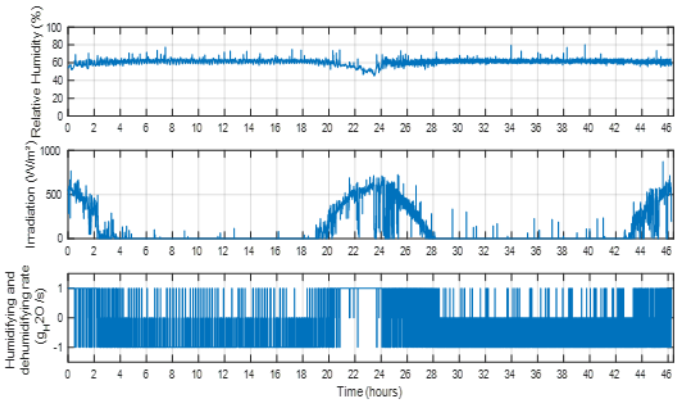


Fig. 10. The Variation of the Relative Humidity with the Controller FLC, the Irradiation Profile and the Humidifying/Dehumidifying Rate.

V. Conclusion

Climate management in greenhouses requires the choice of an automatic control system that is more reliable. In this article, we have developed a greenhouse environmental control system that manages the non-linearity of the system, using the fuzzy model for humidity control and real-time actuators that have been tested on a real greenhouse. This control strategy allows us to effectively regulate greenhouse humidity even in the presence of disturbances.

The simulation results show the performance and efficiency of the developed fuzzy logic control to manage the relative humidity inside the greenhouse in an efficient way. This controller ensures a careful follow-up of the predefined setpoint while reducing the operating time of the actuators.

References

[1] A. Chouchaine, E. Feki, and A. Mami, "Stabilization using a discrete fuzzy PDC controlwith PID controllers and pole placement: application to an experimental greenhouse," J. Control Sci. Eng. pp. 1–9, 2011.

[2] F. Fourati, "Multiple neural control of a greenhouse, Neurocomputing," pp. 138–144, 2014.

[3] M. Taki, Y. Ajabshirchi, S.F. Ranjbar, A. Rohani, and M. Matloobi, "Heat transfer and MLP neural network models to predict inside environment variables and energy lost in a semi-solar greenhouse, Energy Build." pp. 314–329, 2016.

[4] A. Pawlowski, M. Beschi, J.L. Guzmán, A. Visioli, M. Berenguel, and S. Dormido, "Application of SSOD-PI and PI-SSOD event-based controllers to greenhouse climatic control," ISA Trans. Pp. 525–536, 2016.

[5] S. Mohamed, and I.A. Hameed, "A GA-based adaptive neuro-fuzzy controller for greenhouse climate control system," Alex. Eng. J. 2016.

[6] B. Khoshnevisan, S. Rafiee, M. Omid, H. Mousazadeh, and S. Clark, "Environmental impact assessment of tomato and cucumber cultivation

in greenhouses using life cycle assessment and adaptive neuro-fuzzy inference system," J. Clean. Prod. pp. 183–192, 2014.

[7] E. Lachouri, K. Mansouri, M. M. Lafifi, and A. Belmeguenai, "Adaptive neuro-fuzzy inference systems for modeling greenhouse climate," Vol. 7, University Badji Mokhtar Algeria, 2016.

[8] C. E. Lachouri, K. Mansouri, and A. Belmeguenai, "FPGA Implementation of Adaptive Neuro-Fuzzy Inference Systems Controller for Greenhouse Climate," Vol. 7, 2016.

[9] A. Hasni, R. Taibi, B. Draoui, and T. Boulard, "Optimization of greenhouse climate model parameters using particle swarm optimization and genetic algorithms," Energy Procedia 6, pp. 371–380, 2011.

[10] P.J.M. van Beveren, J. Bontsema, G. van Straten, and E.J. van Henten, "Optimal control of greenhouse climate using minimal energy and grower defined bounds," Appl. Energy, pp. 509–519, 2015.

[11] F. Hahn, "Fuzzy controller decreases tomato cracking in greenhouses." Comput. Elect. Agric. pp. 21–27, 2011.

[12] D.M. Atia, and H.T. El-madany, "Analysis and design of greenhouse temperature control usingadaptive neuro-fuzzy inference system," J. Electr. Syst. Inf. Technol. 2016.

[13] S. Revathi, and N. Sivakumaran, "Fuzzy based temperature control of greenhouse," IFACPap. OnLine 49, pp. 549–554, 2016.

[14] F. Hahn,"Irrigation fuzzy controller reduce tomato cracking." Universidad Autonoma Chapingo, México, 2011.

[15] F. Lafont and J.F. Balmat, "Optimized fuzzy control of a greenhouse, Fuzzy Sets Syst," 128, pp. 47–59, 2002.

[16] M.A. Márquez-Vera, J.C. Ramos-Fernández, L.F. Cerecero-Natale, F. Lafont, J.- F. Balmat and J.I. Esparza-Villanueva, "Temperature control in a MISO greenhouse by inverting its fuzzy model," Comput. Electron. Agric. 124, pp. 168–174, 2016.

[17] P. Salgado and J.B. Cunha, "Greenhouse climate hierarchical fuzzy modelling, Control Eng," Pract. 13, pp. 613–628, 2005.

[18] M. Souissi,"Modelisation et commande du climat d'une agricole." Ph.D, these, chapter2, pp. 24-34, 2002.

[19] N. Atyah and H. Afif, "Modeling of greenhouse with PCM energy storage, energy convers." Manag. 49, 3338-3342, 2008.

[20] P. Thirumal, K.S. Amirthagadeswaran and S. Jyabal, "Oprimization of IAQ characteristics of an air-conditioned car using GRA and RSM." J. Mech. Sci. Technol. 28, 1899-1907, 2014.

[21] S. Ashish and S. Tiwari, "Thermal modeling for greenhouse heating by using thermal curtain and earth-air heat exchanger." Build. Environ. 41, 843-850, 2006.

[22] T. H. Nasution and L. A. Harahap, "Predict the Percentage Error of LM35 Temperature Sensor Readings using Simple Linear Regression Analysis," 2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), Medan, Indonesia, 2020, pp. 242-245, doi: 10.1109/ELTICOM50775.2020.9230472.

[23] M. Jomaa, F. Tadeo and A. Mami, "Modeling and experimental validation of the agricultural greenhouse," 2019 IEEE 19th Mediterranean Microwave Symposium (MMS), Hammamet, Tunisia, 2019, pp. 1-5, doi: 10.1109/MMS48040.2019.9157308.

[24] J. Yan, X. Liao, D. Yan and Y. Chen, "Review of Micro Thermoelectric Generator," in Journal of Microelectromechanical Systems, vol. 27, no. 1, pp. 1-18, Feb. 2018, doi: 10.1109/JMEMS.2017.2782748.

[25] N. Lekbangpong, J. Muangprathub, T. Srisawat and A. Wanichsombat, "Precise Automation and Analysis of Environmental Factor Effecting on Growth of St. John's Wort," in IEEE Access, vol. 7, pp. 112848-112858, 2019, doi: 10.1109/ACCESS.2019.2934743.

Selection of Social Media Applications for Ubiquitous Learning using Fuzzy TOPSIS

Caitlin Sam[1], Nalindren Naicker[2], Mogiveny Rajkoomar[3]
Department of Information Systems, Durban University of Technology, Durban, KwaZulu-Natal

***Abstract*—The exponential advancements in Information and Communications Technology has led to its prevalence in education, especially with the arrival of COVID-19. Ubiquitous learning (u-learning) is everyday learning that happens irrespective of time and place and it is enabled by m-learning, e-learning, and social computing such as social media. Due to its popularity, there has been an expansion of social media applications for u-learning. The aim of this research paper was to establish the most relevant social media applications for u-learning in schools. Data was collected from 260 respondents, which comprised learners, and instructors in high schools who were asked to rank 14 of the top social media applications for ubiquitous learning. Fuzzy TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) was the method employed for the ranking of the 14 of the most popular social media applications using 15 education requirements, 15 technology criteria, and 260 decision makers. The simulation was implemented on MATLAB R2020a. The results showed that YouTube was the most likely social media application to be selected for u-learning with a closeness coefficient of 0.9188 and that Viber was the least likely selected social media application with a closeness coefficient of 0.0165. The inferences of this research study will advise researchers in the intelligent decision support systems field to reduce the time and effort made by instructors and learners to select the most beneficial social media application for u-learning.**

Keywords*—*Social media applications; Ubiquitous learning; fuzzy Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS); Multiple criteria decision-making tools

I. Introduction

Ubiquitous learning (u-learning) is a learning paradigm which advocates the employment of ubiquitous computing devices to retrieve learning and teaching resources via. wireless networks [1]. It is evident from the extant literature that education has been indelibly transformed by technology. It enhances and simulates the traditional learning experience, developing instructor and learner behaviour, and creates opportunities for discovery and experimentation [2-4]. As technologies and online tools have progressed, social media has become an essential tool for facilitating applied learning activities [2]. Social media relates to computer-based technology that enables the dissemination of information, thoughts, and ideas through the construction of virtual communities and networks [5]. Social media technologies offer instructors a means to involve learners with self-regulated and time-on-task learning [2, 6, 7]. Generally, social media affords avenues for end-users to communicate and to cultivate connections regardless of geographical barriers and time constraints [2]. In South Africa (SA) alone, out of the 36,4 million people that are internet users, 22 million have social media accounts [8].

School-based learners are captivated by social media [9]. Studies have explained that the adolescent brain has a more sensitive nucleus accumbens and this makes their brain's reward circuitry more activated by social media. Research has also found that the region of the brain linked to visual attention and the social brain are stimulated by social media [10]. An interesting observation is that youth between the ages thirteen and eighteen comprise no less than a fifth of the twelve million Facebook users in SA [9, 11]. A survey revealed that individuals between the ages twelve to nineteen spend ordinary 16,6 hours each week on social networking websites [12]. Recently, there has been a drive towards u-learning with more and more schools being fixated on student centered learning. U-learning is a popular platform to support student centered learning [13]. Furthermore, in unprecedented times such as Covid-19 u-learning is considered as the platform normal teaching and learning practice.

Therefore, there exists a need to support school-based decision makers on which social media applications are most appropriate for u-learning. Much scope exists to address this research gap. A novel application of fuzzy TOPSIS (Technique for Order Preference by Similarity to Ideal Situation) will be used to support decision making on the selection of social media applications for u-learning.

Section 2 discusses the literature review of the top social media applications employed for u-learning, the main technology criteria and education requirements for ranking social media applications for u-learning, and introduces the multi-criteria decision-making tool fuzzy TOPSIS. Section 3 presents the materials and methods employed, while Section 4 discusses the results of fuzzy TOPSIS generated. Section 5 concludes the study.

II. Literature Review

The various features of social media have given rise to a myriad of platforms which can be classified as media sharing tools, social networking tools, experience and resource sharing tools and communication tools [14].

Learners of today have a lot of experience with photo and video-sharing media like YouTube [15, 16]. YouTube facilitates a constructivist classroom which incorporates learning tools for learners actively produce their own learning encounters through creating and viewing videos; and educators can employ these learning tools to meaningfully engage learners [15, 16].

Facebook has collaborative and social characteristics that encourage active participation and social networking of teachers and learners [15]. Facebook has over one billion users and it permits users to share articles, pictures, videos, music, videos, and users' opinions and thoughts. The use of Facebook as a teaching and learning tool has steadily increased over the years [15].

WhatsApp is a cross-application instant messaging service with one and a half billion active users globally [17, 18]. WhatsApp enables end-users to share image, text, video, and voice messages, and to make voice calls and video calls. The differentiation and ubiquity of WhatsApp, has interested a host of studies in various educational research areas [17, 19]. Academic advantages comprise teacher availability, better accessibility of learning materials, and learning extension beyond class hours, peer collaboration, and peer assessment [17, 20]. As an assessment tool, WhatsApp can maintain the anonymity of the learner in the group chat whilst allowing the teacher to read all responses [17].

Facebook Messenger is standalone instant messaging application which was originally part of the Facebook Chat. Facebook Messenger's use in education has rarely been documented but there are several features of the application that can be used in education such as generating attractive posts with videos, text, and images; sharing web content; generating and sharing infographics; posting social proof of learner's success stories to engage prospects; organising contests to entice learners to engage with content; and inspiring an online learning community [21].

WeChat is a text communication and voice messaging service application. The study with Shi and Luo [22] discussed the WeChat teaching platform which facilitated the communication between teachers and learners. Such communication made ubiquitous learning feasible for university students. Furthermore, the study proved that the WeChat teaching platform efficiently developed students' translation competence and facilitated communication in translation teaching [23]. Instagram allows users to edit images and videos with digital filters and asynchronously share and publish images and videos. Additionally, Instagram affords Instagram Stories which publishes time-limited content [24]. The key educational application of Instagram is sharing images or videos for analysis or reference by learners. Other educational affordances of Instagram are supporting direct communication between learners and teachers, facilitating communication, promoting collaboration, posting relevant videos and articles to improve the learning experience [24]. Studies into Instagram for education zone into aspects such as its employment in library contexts and in health and medicine [24].

TikTok has over 1 billion users with most users being between the 14 to 30 age group. Due to the application being very popular in India, EduTok was launched to assist Mathematics and English teaching. Pedagogical affordances of TikTok include motivational influence, delivery of realistic experiences, control and review of content, and engagement of learners as creators [25].

QQ is a powerful communication tool that has proved to support numerous learners with online learning, ensure timeous learner feedback, and active interaction between teachers and learners. The QQ group video allows for synchronous and asynchronous teaching where learners can respond on the video call. Teachers can share their computer screen with learners, which enables learners to learn quickly and conveniently [26].

QZone is a multimedia (audio, image, video, text, etc.) weblog fused with instant message software. It has a user-friendly interface and allows resource access with sharing needs permission. The use of Qzone in the teaching and learning of English has been reported in many studies where the application motivated peer feedback, fulfilled instructional feedback, stimulated in-depth communication, and accelerated learning resource sharing [27-29]. Qzone has no limitation of storage, words, and length, and fulfils the requirements of text editing and processing for various specific purposes [30].

Reddit is the most widespread online content aggregator in the world. Users can publish content, down-vote or up-vote content they dislike or like, and comment on posts. Reddit uses a ranking algorithm to make the most up-voted content more visible in the list of posts [31]. Educational stakeholders can contribute to reflective, meaningful, and stimulating discussions about research, educational policy, politics, research, and technology. Reddit offers teachers and learners a practical platform to engage with educational content in a way that is inquiry-based, open to numerous strategies and engaging [31, 32].

Snapchat's main demographic comprises of millennials. Snapchat is a multimedia sharing and mobile photo messaging. Users can generate Snapchat Stories merging text and visual elements making the application attractive for literacy purposes and multimodal composition [17, 33].

Twitter is a microblogging social media application, which has been found to promote collaborative learning and participation hence transcending traditional classrooms in various studies [34, 35]. Studies also reflected that Twitter was most used for assessment and communication purposes. The most beneficial uses of Twitter included teachers sending homework assignments, test deadlines, and important course information, and facilitating peer interaction [34].

Pinterest is an application for organising, harvesting, sharing, and re-sharing images with comments or updates through republishing. The pedagogical value of Pinterest mostly depends on searching for, organizing, and incorporating digital sources into projects [17, 36]. From a teacher's perspective, Pinterest provides an opportunity for the creation and sense making of instructional resources [17, 37].

Viber is a Voice over IP and an instant messaging application. Users can exchange videos, images, and audio media messages. The study by Farahmand [38] revealed that Viber provided an attractive environment for learners, enhanced communication, improved human interaction, and improved learning [38].

Each apposite social media application for u-learning has numerous characteristics and features, which can be largely

considered as education requirements and technology criteria that need to be investigated by the instructor for it to correlate with the outcomes of the lesson [1]. According to literature, general technology criteria of u-learning include: scalability to accommodate various class sizes, ease of use, the technical support and support availability and hypermediality; cost of use and required equipment [39]; user-focused participation and accessibility standards [40], embedding or integration within an Learning Management System, computer operating system and browser and need for additional downloads; offline access, mobile access, and mobile functionality; privacy, data protection and rights [41]. The typical education requirements of ubiquitous learning that need to be considered are collaboration via. synchronous and asynchronous opportunities, user accountability and diffusion relating to the users' comfort with the tool; teacher facilitation, learning customisation and learning analytics; metacognitive engagement, higher order thinking, and enhancement of cognitive tasks [41]; instructor/learner attitude and beliefs, instructor/learner motivation and incentive, and alignment with learning outcomes and objectives (usefulness) [42].

Therefore, selecting social media applications for u-learning would require a multi-criteria decision-making tool. Existing literature has proved the efficiency of using fuzzy TOPSIS for selecting social media applications with multiple criteria [42, 43]. Fuzzy TOPSIS is a technique first developed by Hwang and Yoon [44], which is employed to systematically and objectively evaluate various alternatives against multiple selected criteria to solve multi-criteria decision-making (MCDM) problems [45]. The rationale of fuzzy TOPSIS is that the solution that minimises the cost criteria and maximises the benefit criteria will be denoted by an alternative with the smallest distance from the positive ideal solution (PIS). Alternately, the solution that minimises the benefit criteria and maximises the cost criteria will be denoted by an alternative with the largest distance from the negative ideal solution (NIS) [46]. Since human opinions are vague and cannot be quantified, classical TOPSIS cannot be applied as it assigns crisp numerical data to the alternative's performance ratings and criteria weights [47]. Thus, the fuzzy set theory has been incorporated in many MCDM approaches, namely, TOPSIS [48].

In recent times, fuzzy TOPSIS techniques and its functions have been covered extensively by more scholars [49]. This paper will use fuzzy TOPSIS to select social media applications for u-learning in high schools.

III. Research Method

A. Population

The target population was school-based instructors and learners from the eThekwini Region, namely, Pinetown District and Umlazi District in KwaZulu Natal, South Africa. As per the March 2020 Schools Masterlist Data derived from the Department of Basic Education in South Africa [50] the study's population size was approximately 129 421 individuals which comprised 4 853 school-based instructors and 124 388 learners. For clarification of the target population, the school-based instructors and learners were treated as a single population. Thus, the decision makers were inclusively the learners and instructors that used social media applications for u-learning.

B. Sample

In accordance with the guidelines set by Sekaran and Bougie [51], as the target population size of the study was approximately 129 421 (Department of Basic Education, 2020), the sample size was 384 respondents. However, the response rate was 67,71% with the total number of responses received being 260. According to literature, the goal of researchers conducting a survey questionnaire should be approximately 60% [52]. Thus, the response rate of the current study was acceptable.

C. Questionnaire

To collect the dataset, a link to the survey questionnaire was forwarded to respondents' devices. Online survey questionnaires on Google Forms delivered a user-friendly interface on all respondents' devices and a cost-effective and time-efficient data collection. When compared to other survey-generating platforms, an unlimited number of matrix-formatted questions could be generated on Google Forms. The questionnaire comprised closed ended questions using a Likert scale where the respondent chose one suitable answer for each question. The possible answers were 'very poor', 'poor', 'fair', 'good', or 'very good' and 'not sure'. The 'not sure' option was presented to respondents who were unfamiliar with certain social media applications. The first part of the questionnaire consisted of questions pertaining to the respondents' demographic data. The questions involved education requirements and technology criteria resulting from the existing literature that is important to the social media diffusion management in school-based u-learning. The 30 most commonly occurring education requirements (15) and technology criteria (15) were chosen. Table I shows a snippet of the survey questionnaire.

ADAPTABILITY: the social media application provides personalized learning, which aims to give efficient, effective, and customized learning paths to attract each learner.

TABLE I. Snippet of Survey Questionnaire Generated on Google Forms

	Very Good	Good	Fair	Poor	Very Poor	Not Sure
Facebook	❑	❑	❑	❑	❑	❑
WhatsApp	❑	❑	❑	❑	❑	❑
YouTube	❑	❑	❑	❑	❑	❑
Facebook Messenger	❑	❑	❑	❑	❑	❑
Instagram	❑	❑	❑	❑	❑	❑
TikTok	❑	❑	❑	❑	❑	❑
QQ	❑	❑	❑	❑	❑	❑
QZone	❑	❑	❑	❑	❑	❑
Reddit	❑	❑	❑	❑	❑	❑
Pinterest	❑	❑	❑	❑	❑	❑
WeChat	❑	❑	❑	❑	❑	❑
SnapChat	❑	❑	❑	❑	❑	❑
Twitter	❑	❑	❑	❑	❑	❑
Viber	❑	❑	❑	❑	❑	❑

The data was analysed and synthesised using fuzzy TOPSIS programmatically on MATLAB R2020a.

D. Fuzzy TOPSIS Method

The fuzzy TOPSIS method assesses various alternatives against the selected criteria. In the TOPSIS method, the best alternative is typified by a calculated distance that is closest to the Fuzzy Positive Ideal Solution (FPIS) and furthest from the Fuzzy Negative Ideal Solution (FNIS) [53]. An FPIS involves alternatives with the best performance values and the FNIS involves alternatives with the worst performance values. In fuzzy TOPSIS linguistic variables are utilised to describe all the ratings and weights which in turn are expressed by fuzzy numbers [54]. The fuzzy set theory illustrates a triangular fuzzy number (TFN) as a triplet (a; b; c) where:

$F(x)$ can be expressed as:

$$\mu_{\tilde{A}}(X) = \begin{cases} \frac{x-a}{b-a}, a \le x \le b \\ \frac{c-x}{c-b}, b \le x \le c \\ 0, otherwise \end{cases} \quad (1)$$

The fuzzy triplets a, b and c are real numbers and $a < b < c$ [53]. There are a series of steps as outlined below that must be adhered to when conducting the fuzzy TOPSIS algorithm [54].

E. Steps of the Fuzzy TOPSIS method

The steps below illustrate the fuzzy rating and importance weight of the k^{th} decision maker, about the i^{th} alternative on j^{th} criterion [53].

Step 1: Assignment of ratings to the alternatives and the criteria.

Assume that there are m possible alternatives called $A = \{A_1; A_2; \dots; A_m\}$ which must be evaluated against n criteria, $C = \{C_1; C_2; \dots; C_n\}$. Criteria weights are represented by $w_g\ (g = 1,2, \dots, n)$. Each decision maker's $D_k(k = 1,2, \dots, k)$ ratings for each alternative $A_i(i = 1,2, \dots, m)$ regarding criteria $C_j(j = 1,2, \dots, n)$ are indicated by $\tilde{R}_k = \tilde{x}_{igk}$ $\tilde{R}$ with the membership function $\mu_{\tilde{R}_k}(x)$.

The fuzzy ratings for the criteria by decision makers are shown in the Table II.

The fuzzy ratings for the alternatives are shown in the Table III.

Step 2: Computation of aggregated fuzzy ratings for the alternatives and the criteria.

- If the importance weight and fuzzy rating of the decision maker k are [53]:

$$\tilde{w}_{igk} = (w_{gk1}, w_{gk2}, w_{gk3}) \quad (2)$$

and $\tilde{x}_{igk} = (a_{igk}, b_{igk}, c_{igk})$,

$$i = 1,2, \dots m\ , g = 1,2, \dots, n \quad (3)$$

respectively, then the aggregated fuzzy ratings $\tilde{x}_{ijk}$ of each alternative relating to each criterion is presented as:

$$\tilde{x}_{ijk} = (a_{ig}, b_{ig}, c_{ig}) \quad (4)$$

where: $a_{ig} = min_k\{a_{igk}\}$ (5)

$$b_{ig} = \frac{1}{k}\sum_{k=1}^{k} b_{igk} \quad (6)$$

$$c_{ig} = max_k\{c_{igk}\} \quad (7)$$

- The aggregated fuzzy weights $(\tilde{w}_{ij})$ of each criterion is calculated as

$$\tilde{w}_{ig} = (w_{g1}, w_{g2}, w_{g3}) \quad (8)$$

where: $w_{g1} = min_k\{w_{gk1}\}$, $w_{g2} = \frac{1}{k}\sum_{k=1}^{k} w_{jk2}$,

$$w_{g3} = max_k\{w_{jk3}\} \quad (9)$$

Step 3: Computation and normalisation of the fuzzy decision matrix.

The fuzzy decision matrix is computed as such:

$$\tilde{D} = \begin{matrix} A_1 \\ A_2 \\ A_m \end{matrix} \begin{pmatrix} \tilde{x}_{11} & \tilde{x}_{12} \dots & \tilde{x}_{1n} \\ \tilde{x}_{21} & \tilde{x}_{22} \dots & \tilde{x}_{2n} \\ \tilde{x}_{m1} & \tilde{x}_{m2} \dots & \tilde{x}_{mn} \end{pmatrix} \quad (10)$$

$$\tilde{W} = (\tilde{w}_1, \tilde{w}_2\ , \dots, \tilde{w}_n) \quad (11)$$

The various criteria scales are normalised into a comparable scale. The normalised fuzzy decision matrix is $\tilde{\boldsymbol{R}}$ and is presented as:

$$\tilde{\boldsymbol{R}} = [\tilde{\boldsymbol{r}}_{\boldsymbol{ig}}]\ \boldsymbol{m} \times \boldsymbol{n}\ , \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{m};\ \boldsymbol{g} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{n} \quad (12)$$

thus, the benefit criteria: $\bar{\boldsymbol{r}}_{\boldsymbol{ig}} = \frac{\boldsymbol{a_{ig}}}{\boldsymbol{c_g^+}}, \frac{\boldsymbol{b_{ig}}}{\boldsymbol{c_g^+}}, \frac{\boldsymbol{c_{ig}}}{\boldsymbol{c_g^+}}$ (13)

where: $\boldsymbol{c_g^+} = \boldsymbol{max_i c_{ig}}$ of the benefit criteria (14)

and for cost criteria: $\tilde{\boldsymbol{r}}_{\boldsymbol{ig}} = \frac{\bar{\boldsymbol{a}}_{\boldsymbol{j}}}{\boldsymbol{c_{ij}}}, \frac{\bar{\boldsymbol{a}}_{\boldsymbol{j}}}{\boldsymbol{b_{ij}}}, \frac{\bar{\boldsymbol{a}}_{\boldsymbol{j}}}{\boldsymbol{a_{ij}}}$ (15)

where: $\bar{\boldsymbol{a}}_{\boldsymbol{j}} = \boldsymbol{minimim_i a_{ig}}$ of the cost criteria (16)

TABLE II. LINGUISTIC TERMS FOR CRITERIA

Linguistic term	Triangular fuzzy numbers
Very poor (VP)	(1, 1, 3)
Poor (P)	(1, 3, 5)
Fair (F)	(3, 5, 7)
Good (G)	(5, 7, 9)
Very good (VG)	(7, 9, 9)

TABLE III. LINGUISTIC TERMS FOR ALTERNATIVES

Linguistic term	Triangular fuzzy numbers
Very poor (VP)	(1, 1, 3)
Poor (P)	(1, 3, 5)
Fair (F)	(3, 5, 7)
Good (G)	(5, 7, 9)
Very good (VG)	(7, 9, 9)

Step 4: Computation of the weighted normalised matrix

The weighted normalised fuzzy decision matrix $\tilde{\boldsymbol{v}}_{\boldsymbol{ij}}$ is derived by:

$$\tilde{\boldsymbol{V}} = \left[\tilde{\boldsymbol{v}}_{\boldsymbol{ij}}\right]_{\boldsymbol{m}\times\boldsymbol{n}} \; \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{n}; \; \boldsymbol{j} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{m} \tag{17}$$

where: $\tilde{\boldsymbol{V}}_{\boldsymbol{ij}} = \tilde{\boldsymbol{r}}_{\boldsymbol{ij}} \times \tilde{\boldsymbol{w}}_{\boldsymbol{j}}$ (18)

Step 5: Computation of the FPIS ($\boldsymbol{A}^{+}$) and FNIS ($\boldsymbol{A}^{-}$)

The FPIS of the alternatives are computed as:

$$\boldsymbol{A}^{+} = (\tilde{\boldsymbol{v}}_{\boldsymbol{1}}^{+}, \tilde{\boldsymbol{v}}_{\boldsymbol{2}}^{+}, \dots, \tilde{\boldsymbol{v}}_{\boldsymbol{m}}^{+}) \tag{19}$$

where: $\boldsymbol{v}_{\boldsymbol{i}}^{+} = \boldsymbol{max}_{\boldsymbol{i}}\{\boldsymbol{v}_{\boldsymbol{ig3}}\}$,

$$\boldsymbol{g} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{n}; \; \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{m} \tag{20}$$

and the FNIS of the alternatives are computed as:

$$A^{-} = (\tilde{v}_{1}^{-}, \tilde{v}_{2}^{-}, \dots, \tilde{v}_{m}^{-}) \tag{21}$$

where: $\tilde{v}_{j}^{-} = min_{i}\{v_{ig1}\}$,

$$g = 1, 2, \dots, n; \; i = 1, 2, \dots, m \tag{22}$$

Step 6: Calculation of the distance of each alternative to get theA^{+} matrix and A^{-} matrix

The distance (d^{+}, d^{-}) of the A^{+} and the A^{-} from each weighted alternative $i = 1,2, \dots, m$ is calculated using formula:

$$d(\bar{a}\bar{b}) = \sqrt{\frac{1}{3}\left[(a-a')^{2}+(b-b')^{2}+(c-c')^{2}\right]} \tag{23}$$

Step 7: Calculating the distance of each weighted alternative

The sum of the distance of each weighted alternative is calculated by:

$$d_{i}^{+} = \sum_{i=1}^{m} d_{v}(\tilde{v}_{ig}, \tilde{v}_{ig}^{+}), i = 1,2, \dots, m;$$

$$g = 1,2, \dots, n \tag{24}$$

$$d_{i}^{-} = \sum_{i=1}^{m} d_{v}(\tilde{v}_{ig}, \tilde{v}_{ig}^{-}), i = 1,2, \dots, m; \; g = 1,2, \dots, n \tag{25}$$

Step 8: Computation of the closeness coefficient (CC_{i}) of each alternative

CC_{i} signifies the distances to the FNIS and the FPIS, simultaneously. The calculation of each alternative's CC_{i} is as follows:

$$(CC_{i}) = \frac{d_{1}^{-}}{d_{1}^{-}+d_{1}^{+}}, \; i = 1, 2, \dots, m \tag{26}$$

Step 9: Ranking the alternatives

The ranking order of all the alternatives can be derived from the CC_{i}. The closer the CC_{i} is to the FPIS and the farthest it is from the FNIS the better the alternative. This means the higher the CC_{i} the higher the rank of the alternative [54].

F. MATLAB R2020a

Fuzzy TOPSIS was coded using MATLAB R2020a running on a windows i7 computer. MATLAB® merges a desktop environment set for design processes and iterative analysis with a programming language that articulates array and matrix mathematics directly [55]. The MATLAB code followed a modular programming design. Sub programmes for Weighted normalized fuzzy decision matrix for alternatives; Aggregate Fuzzy Weights for Criteria, Distances, and the final calculation of CC_{i} values were called by a main program. The dataset was read by the code using a 'xlsread' statement. Intermediate matrices were generated in the MATLAB workspace. The results obtained is presented in the next section.

IV. Results and Discussion

This section presents the results of the 260 respondents' rating of the social media applications using the criteria for the evaluation of social media applications for u-learning. In this study, criteria comprised the technology criteria and the education requirements. In terms of fuzzy TOPSIS an analysis was conducted using 14 alternatives, and 30 criteria with ratings by 260 decision makers (DM). As the criteria formed the top 30 technology criteria and education requirements, the ratings of the attributes were either Good (G) or Very Good (VG). Thus, their weightings were either (5, 7, 9) or (7, 9, 9), respectively. Table IV reflects the weightage of the criteria (C), namely education requirements and technology criteria:

Table IV shows 15 education requirements and 15 technology criteria for selection of social media applications for u-learning. The weightage for each criteria is given in terms of a triangular fuzzy number. The criteria weights will be factored into decision making for the ranking of social media applications for u-learning.

The social media applications were the alternatives in the study and were labelled as per Table V for the simulation on MATLAB.

The multi-criteria decision making using Fuzzy TOPSIS is applied to rank the set of 14 alternatives as shown in Table V.

Table VI shows an extract of the combined decision matrix for A1 which was Facebook. The decision makers rated the alternatives in terms of linguistic scales.

The linguistic terms were assigned fuzzy numbers. The "Not Sure" option termed "N" was zero-rated as it is not a linguistic term identifiable on fuzzy TOPSIS. The aggregated fuzzy ratings for the alternatives and the criteria were computed. Once combined decision matrix was normalised and multiplied by the criteria weightage, the weighted normalised fuzzy decision matrix was achieved. The FPIS (A^{+}) and FNIS (A^{-}) were determined and the distance of each alternative to get the A^{+} and A^{-} matrix was calculated. Table VII shows the A^{+} and A^{-} matrix of the Facebook alternative.

The sum of the distance of each weighted alternative was calculated using formula (24) and (25). Thereafter the CC_{i} was calculated using formula (26). Table VIII reflects the CC_{i} and ranking of each alternative.

TABLE IV. WEIGHTED ATTRIBUTES OF SOCIAL MEDIA APPLICATIONS FOR UBIQUITOUS LEARNING

C#	Education Requirements	Weightage	C#	Technology Criteria	Weightage
1.	Ownership of Learning	(5, 7, 9)	16.	Operational Stability	(7, 9, 9)
2.	Adaptability	(5, 7, 9)	17.	Fault Tolerance of Technology	(7, 9, 9)
3.	Quality Assurance	(5, 7, 9)	18.	Hypermediality	(7, 9, 9)
4.	Peer Learning	(5, 7, 9)	19.	Multimedia Control	(7, 9, 9)
5.	Instructional Design	(5, 7, 9)	20.	Security of Technology	(7, 9, 9)
6.	Academic Integrity	(5, 7, 9)	21.	Facilitation of e-Content	(7, 9, 9)
7.	U-learning training factors	(5, 7, 9)	22.	Technical Information	(7, 9, 9)
8.	Archiving/ Repository	(5, 7, 9)	23.	Software Characteristics Quality	(5, 7, 9)
9.	Social Interaction	(5, 7, 9)	24.	Ease of Use	(7, 9, 9)
10.	Curriculum Management	(7, 9, 9)	25.	Operating System	(5, 7, 9)
11.	Facilitation	(5, 7, 9)	26.	Browser	(7, 9, 9)
12.	Learning Analytics	(5, 7, 9)	27.	Access on a Mobile Platform	(5, 7, 9)
13.	Enhancement of Cognitive Tasks	(5, 7, 9)	28.	Data Privacy and Ownership	(7, 9, 9)
14.	Higher Order Thinking	(5, 7, 9)	29.	Downloading, Saving and Exporting Data	(5, 7, 9)
15.	Metacognitive Engagement	(5, 7, 9)	30.	Additional Download	(5, 7, 9)

TABLE V. LABELLED SOCIAL MEDIA APPLICATIONS

Social Media Application	Alternative (A)	Social Media Application	Alternative (A)
Facebook	1	QZone	8
WhatsApp	2	Reddit	9
YouTube	3	Pinterest	10
Facebook Messenger	4	WeChat	11
Instagram	5	SnapChat	12
TikTok	6	Twitter	13
QQ	7	Viber	14

TABLE VI. ASSIGNMENT OF RATINGS TO FACEBOOK BY DECISION MAKERS

Criteria	FACEBOOK (A1)								
	DM1	DM2	DM3	DM4	DM5	DM6	DM7	DM8…	DM260
C1	N	P	N	N	N	F	G	G	G
C2	F	P	VP	N	N	G	F	G	G
C3	F	P	N	N	N	F	F	F	VG
C4	N	P	N	N	N	F	G	G	F
C5	N	P	F	N	N	F	VG	G	F
C6	F	P	N	N	N	G	F	G	VG
C7	N	P	N	N	N	F	N	G	G
C8	N	P	F	VP	VP	G	VG	VG	VG
C9	N	P	N	N	N	F	G	G	F
C10…	N	N	F	N	N	G	G	VG	F
C26	G	N	VG	N	VP	G	VG	G	F
C27	G	N	VP	N	N	VG	VG	G	G
C28	G	N	G	N	F	VG	VG	VG	VG
C29	G	N	N	N	N	VG	VG	VG	VG
C30	G	N	N	N	N	VG	VG	VG	VG

TABLE VII. THE FPIS (A^+) AND FNIS (A^-) OF FACEBOOK

Criteria	FPIS (A^+)	FNIS (A^-)	Criteria	FPIS (A^+)	FNIS (A^-)
C1	1.4887	0.5526	C16	1.2538	0.5993
C2	1.5578	0.5285	C17	0.7616	0.4110
C3	1.4645	0.5008	C18	0.9930	1.0621
C4	1.2866	0.8221	C19	0.8342	1.0293
C5	1.5405	0.5077	C20	0.6425	0.6597
C6	1.0949	0.4680	C21	1.0621	0.6425
C7	1.4300	0.6701	C22	0.7236	0.6459
C8	1.9274	0.0000	C23	0.7461	0.6822
C9	1.4922	0.8894	C24	0.9067	0.9153
C10	1.4783	0.0846	C25	0.6580	0.9205
C11	1.0138	0.5526	C26	0.7443	0.8894
C12	0.9032	0.5665	C27	0.4473	0.9792
C13	1.1312	0.6079	C28	0.4628	0.6908
C14	1.3246	0.4197	C29	0.7098	0.8583
C15	0.8894	0.5526	C30	0.6217	0.7910

TABLE VIII. CLOSENESS COEFFICIENT (CC_i) AND RANKING OF ALTERNATIVES

Ranking #	Alternative	(CC_i)
1	YouTube	0.9188
2	WhatsApp	0.8691
3	Instagram	0.4835
4	TikTok	0.3877
5	Facebook	0.3817
6	Facebook Messenger	0.3249
7	Pinterest	0.2801
8	SnapChat	0.2484
9	Twitter	0.2108
10	Reddit	0.1601
11	WeChat	0.1379
12	QQ	0.0720
13	QZone	0.0344
14	Viber	0.0165

Since $CC_{YouTube} > CC_{WhatsApp} > CC_{Instagram} > CC_{TikTok} > CC_{Facebook} \dots$, YouTube is the preferred social media platform considering the given criteria. For example, if the criterion 'Ownership of Learning' is explored in relation to YouTube, it is evident that YouTube fulfils all the expectations of the end-user being motivated, engaged and self-directed. According to Husain *et al*. [58], YouTube allows teachers and learners ubiquitous access on any digital device to videos and content made by subject experts from all around the world that would have otherwise been expensive to acquire. Teachers and learners can learn skills from step-by-step videos made by skilled individuals on YouTube that can be replayed countless of times. YouTube is an unlimited digital library with multimedia tools and is a platform where teachers and learners can make their own videos to display their skills and talents on a global scale. The use of YouTube facilitates for flipped classrooms which allows individuals to take ownership of their learning in a manner that is free and fair to all that want to learn [59]. Regarding Viber, several sources revealed that it has poor messaging services and voice call quality without Wi-Fi connection [60].

V. CONCLUSION

Social media has become an integral tool in the affordance of ubiquitous learning for schools, especially in unprecedented times such as COVID-19. According to HelloYes [56], the most used social media applications in South Africa in ranking order are WhatsApp, YouTube, Facebook, FB Messenger, Instagram, Twitter, Pinterest, LinkedIn, Snapchat, Skype, Reddit, TikTok, Tumbler, WeChat, Twitch, and Viber. However, according to Smart Insights [57] Generation Z individuals prefer the following social media applications:

Facebook, YouTube, WhatsApp, Facebook Messenger, WeChat, Instagram, TikTok, QQ, QZone, Reddit, Snapchat, Twitter, Pinterest, and Viber. This study ranked 14 of the top social media applications at the time using the multiple criteria decision-making tool fuzzy TOPSIS and 30 ubiquitous learning criteria with the 260 decision makers. The results revealed that YouTube proved to be the most significant social media application for ubiquitous learning with a $CC_i = 0{,}9188$, closely followed by WhatsApp with a $CC_i = 0{,}8691$. Viber was identified as being the least likely application suitable to ubiquitous learning with a $CC_i = 0{,}0165$.

A limitation of the current study was that it had to be representative of the high school population in the Pinetown and Umlazi Districts in the province of KwaZulu Natal, South Africa, responses from a large sample population had to be acquired. The opinions of 260 decision makers comprising instructors and learners in high schools was a novel approach that proved to be initially challenging to configure on MATLAB R2020a. An implication of this study is that it contributed to the gap of literature on the use of fuzzy TOPSIS in school-based education and enriched the literature on the application of fuzzy TOPSIS within a South African schooling context. Additionally, given the current pandemic, getting "online now" presents a different focus for decision making around social media platforms as ubiquitous learning tools. Future studies will focus on intelligent decision support systems to reduce the time and effort made by instructors and learners to select the most beneficial social media application for ubiquitous learning.

ACKNOWLEDGMENTS

Kind acknowledgement to the Durban University of Technology for making funding opportunities and resources available for this research project.

REFERENCES

[1] C., Sam, N. Naicker, and M. Rajkoomar, Meta-analysis of artificial intelligence works in ubiquitous learning environments and technologies. International Journal of Advanced Computer Science and Applications. Vol. (11), No. (9), 2020, p. 603-613.

[2] J., Purvis, H.M. Rodger, and S. Beckingham, Experiences and perspectives of social media in learning and teaching in higher education. International Journal of Educational Research Open. Vol. (1), 2020, p. 1-17.

[3] M., Chen, Students' perceptions of the educational usage of a Facebook group. Journal of Teaching in Travel and Tourism. Vol. (18), No. (4), 2018, p. 332-348.

[4] D.R., Garrison and H. Kanuka, Blended learning: Uncovering its transformative potential in higher education. Internet and Higher Education. Vol. (7), No. (2), 2004, p. 95-105.

[5] C., Greenhow, S.M. Galvin, and K.B.S. Willet, What should be the role of social media in education? Behavioural and Brain Sciences. Vol. (6), No (2), 2019, p. 178-185.

[6] A. J., Purvis, H.M. Rodger, and S. Beckingham, Engagement or distraction: The use of social media for learning in higher education. Student Engagement and Experience Journal. Vol. (5), No (1), 2016, p. 1-5.

[7] N., Dabbagh and A. Kitsantas, Personal learning environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. Internet and Higher education. Vol. (15), No. (1), 2011, p. 3-8.

[8] Statista. Digital population in South Africa, 2020, Accessed on 30 January 2020; Available from: https://www.statista.com/ statistics/685134/south-africa-digital-population.

[9] Shava, H. and W.T. Chinyamurindi, Determinants of social media usage among a sample of rural South African youth. South African Journal of Information Management. Vol. (20), No. (1), 2018, p. 1-8.

[10] S., Wolpert, Teenage brain on social media: Study sheds light on influence of peers and much more. 2016; Accessed on 30 January 2020; Available from: www.sciencedaily.com/releases /2016/05/ 160531104423.htm.

[11] BusinessTech. These are the biggest social media and chat platforms in 2019. 2019; Accessed on 30 January 2020; Available from: https://businesstech.co.za/news/internet/296752/ these-are-the-biggest-social-media-and-chat-platforms-in-2019/.

[12] J., Lu, Q. Hao, and M. Jing, Consuming, sharing, and creating content: How young students use new social media in and outside school. Computers in Human Behaviour. Vol. (64), 2016, p. 55-64.

[13] D., Jaiswal, New approaches in e-learning: E-learning, m-learning and u-learning. Scholarly Research Journal for Interdisciplinary Studies. Vol. (1), No. (2), 2012, p. 197-203.

[14] G.E., Zgheib and N. Dabbagh, Social media learning activities (SMLA): Implications for design. Online Learning. Vol. (24), No. (1), 2020, p. 50-66.

[15] M., Mei, and L. Yeo, Social media and social networking applications for teaching and learning. European Journal of Science and Mathematics Education. Vol (2), No. (1), 2014, p. 53-62.

[16] R., Mullen and L. Wedwick, Avoiding the digital abyss: Getting started in the classroom with youtube, digital stories and blogs. Clearing house: A Journal of Educational Strategies, Issues and Ideas, 2008. Vol. (82), No. (2), 2008, p. 66-69.

[17] S., Manca, Snapping, pinning, liking, or texting: Investigating social media in higher education beyond Facebook. The Internet and Higher Education. Vol. (44), 2020, p. 1-13.

[18] V., Sigurdsson, R.V., Menon, A.G., Hallgrímsson, N.M., Larsen, & A., Fagerstrøm. Factors affecting attitudes and behavioral intentions toward in-app mobile advertisements. Journal of Promotion Management, Vol. (24), No. (5), 2018, pp 694-714.

[19] C., Pimmer and P. Rambe, The inherent tensions of "instant education": A critical review of mobile instant messaging. The International Review of Research in Open and Distance Learning. Vol. (19), No. (5), 2018, p. 218-237.

[20] D., Bouhnik and M. Deshen, WhatsApp goes to school: Mobile instant messaging between teachers and students. Journal of Information Technology Education: Research. Vol. (13), No. (1), 2014, p. 217-231.

[21] P., Smutny and P. Schreiberova, Chatbots for learning: A review of educational chatbots for the Facebook Messenger. Computers and Education. Vol. 151, 2020, p. 1-11.

[22] S., Shi and G. Luo, Application of WeChat teaching platform in interactive translation teaching. International Journal of Emerging Technologies in Learning. Vol. 11, No. (9), 2016, p. 71-75.

[23] M. G., Elyazgi, et al., Evaluating the factors influencing e-book technology acceptance among school children using TOPSIS technique. Journal of Soft Computing and Decision Support Systems. Vol. (3), No. (2), 2016, p. 11-25.

[24] M. N. K., Boulos, D.M. Giustini, and S. Wheeler, Instagram and Whatsapp in health and healthcare: An overview. Future Internet. Vol. (8), No. (37), 2016, p. 1-14.

[25] CommonwealthofLearning. Importance of TikTok type videos for learning, 2020, Accessed on 20 January 2020; Available from: https://www.col.org/news/col-blog/importance-tiktok-type-videos-learning.

[26] G., Nan, Application of QQ Classroom in modern university education. Advances in Higher Education. Vol. 4, No. (5), 2020, p. 57-59.

[27] X.P., Wen and W.B. Lai, The application of Qzone in middle school English teaching. The Teaching of Politics. Vol. (12), 2012, p. 145-146.

[28] X.J., Wang, Using Qzone in the process of implementing English language teaching methodology course. Journal of Hubei University of Education. Vol. 26, No. (7), 2009, p. 114-115.

[29] Q.J., Du, On course design of business English teaching based on QQ platform. Overseas English. Vol. (7), No. (1), 2013, p. 106-108.

[30] G., Xianwei, M., Samuel, and A. Asmawi, Qzone weblog for critical peer feedback to improve business english writing: A case of chinese undergraduates. The Turkish Online Journal of Educational Technology. Vol. 15, No. (3), 2016, p. 131-140.

[31] R.P., Tannebaum, Reddit and the social studies: Exploring the r/democratic curriculum. The Social Science. Vol. (109), No. (3), 2018, p. 167-175.

[32] M.M., Diacopoulos, Untangling Web 2.0: Charting Web 2.0 tools, the NCSS guidelines for effective use of technology, and Bloom's Taxonomy. The Social Studies. Vol. (106), No. (4), 2015, p. 139-148.

[33] J.T., Bartels, Soft(a)ware in the English classroom. English Journal. Vol. (106), No. (5), 2017, p. 90-92.

[34] Y. Tang and K.F. Hew, Using Twitter for education: Beneficial or simply a waste of time? Computers and Education. Vol. (106), No. (1), 2017, p. 97-118.

[35] F., Gao., T. Luo, and K. Zhang, Tweeting for learning: A critical analysis of research on microblogging in education published in 2008-2011. British Journal of Educational Technology. Vol. (43), No. (3), 2012, p. 783-801.

[36] C. Geraths and M. Kennerly, Pinvention: Updating commonplace books for the digital age. Communication Teacher. Vol. (29), No. (3), 2015, p. 116-172.

[37] S., Hu , et al., What do teachers share within socialized knowledge communities: A case of Pinterest. Journal of Professional Capital and Community, 2018. Vol. (3), No. (2), 2018, p. 97-122.

[38] F., Farahmand, The effects of using Viber on Iranian EFL university students' vocabulary learning (an interactionist view). International Journal of Social and Educational Innovation. Vol. (3), No. (5), 2016, p. 31-38.

[39] E., Bagarukayo and B. Kalema, Evaluation of e-learning usage in South African universities: A critical review. International Journal of Education and Development using Information and Communication Technology. Vol. (11), No. (2), 2015, p. 168-183.

[40] G., Grigoraș, D. Dănciulescub, and C. Sitnikovc, Assessment criteria of e-learning environments quality. Procedia Economics and Finance. Vol. (16), 2014, p. 40-46.

[41] L.M. Anstey and G.P.L. Watson, A rubric for evaluating e-learning tools in higher education. EDUCASE Review, 2018: Accessed on 30 January 2020, Available online at https://er.educause. edu/articles/2018/9/a-rubric-for-evaluating-e-learning-tools-in-higher-education.

[42] M., Meyliana, A.N. Hidayanto, and E.K. Budiardjo, Evaluation of social media channel preference for student engagement improvement in universities using entropy and TOPSIS method. Journal of Industrial Engineering and Management. Vol. (8), No. (5), 2015, p. 1676-1697.

[43] A.D.S. Sirait, et al. Evaluation of social media preference as e-participation channel for students using fuzzy AHP and TOPSIS. in 2018 4th International Conference on Computing, Engineering, and Design. 2018. Bangkok, Thailand, Bangkok: Institute of Electrical and Electronics Engineers Inc.

[44] C., Hwang and K. Yoon, Multiple Attribute Decision Making. 1 ed. Methods and Applications A State-of-the-Art Survey. 1981, Berlin Heidelberg: Springer, p 58-91

[45] B., Sodhi and T.V. Prabhakar. A Simplified Description of Fuzzy TOPSIS. 2012; Accessed on 30 January 2020; Available from: http://arxiv.org/abs/1205.5098.

[46] R.K., Singh and L. Benyoucef, A fuzzy TOPSIS based approach for e-sourcing. Engineering Applications of Artificial Intelligence. Vol. (24), No. (3), 2011, p. 437-448.

[47] G., Kannan, S. Pokharel, and P.S. Kumar, A hybrid approach using ISM and fuzzy TOPSIS for the selection of reverse logistics provider. Resources, Conservation and Recycling. Vol. (54), No. (1), 2009, p. 28-36.

[48] A., Kelemenis, K. Ergazakis, and D. Askounis, Support managers' selection using an extension of fuzzy TOPSIS. Expert Systems with Applications. Vol. (38), No. (3), 2011, p. 2774–2782.

[49] K., Ayebi-Arthur, E-learning, resilience and change in higher education: Helping a university cope after natural disaster. E-learning and Digital Media. Vol. (14), No. (1), 2017, p. 259-274.

[50] Department of Basic Education. School masterlist data. 2020; Accessed on 30 January 2020; Available from: https://www.education. gov.za/Programmes/EMIS/ EMISDownloads.aspx.

[51] U ., Sekaran and R.J. Bougie, Research Methods for Business: A Skill Building Approach. 7th ed. 2016, Chichester, West Sussex: John Wiley and Sons.

[52] J.E., Fincham, Response rates and responsiveness for surveys, standards, and the journal. American Journal of Pharmaceutical Education. Vol. (72), No. (2), 2008, p. 1-3.

[53] H., Han and S. Trimi, A fuzzy TOPSIS method for performance evaluation of reverse logistics in social commerce platforms. Expert Systems with Applications. Vol. (103), No. (1), 2018, p. 133-145.

[54] W., Huang, et al., Applying fuzzy technique for order preference by similarity to ideal solution (TOPSIS) in the slection of best candidate: A case study on interview performance. British Journal of Economics, Finance and Management Sciences. Vol. (17), No. (1), 2020, p. 36-49.

[55] MathWorks. MATLAB. 2020; Accessed on 30 January 2020; Available from: https://www.mathworks.com/ products/ matlab.html.

[56] HelloYes. Digital statistics and usage in South Africa 2020. 2020; Accessed on 30 January 2020; Available from: https://www. helloyes.co.za/digital-statistics-and-usage-in-south-africa-in-2020/.

[57] SmartInsights. Global social media research summary. 2020; Accessed on 30 January 2020; Available from: https://www.smartinsights. com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

[58] K., Husain, et al., Benefits of YouTube videos usage in students' learning. Journal of Human Capital Development. Vol. (5), No. (2), 2012, p. 1-8.

[59] Learn From Blogs. Why YouTube is important for teachers and students. 2021; Accessed on 30 January 2020; Available from: https://learnfromblogs.com/why-youtube-is-important-for-teachers-and-students.

[60] A., Chand, and Chand, P., Advantages and disadvantages of using Viber as distance student learning tool in a regional university. International Journal of Instructional Technology and Distance Learning. Vol. (14), No. (12), 2017, p. 54-60.

Nonlinear Rainfall Yearly Prediction based on Autoregressive Artificial Neural Networks Model in Central Jordan using Data Records: 1938-2018

Suhail Sharadqah[1], Soraya Mercedes Perez[5]
Department of Natural Resources and Chemical Engineering
Tafila Technical University, Tafila, Jordan

Mohammad A Obeidat[3]
Department of Electrical Power and Mechatronics Engineering, Tafila Technical University, Tafila, Jordan

Ayman M Mansour[2]
Department of Communication, Electronics and Computer Engineering, Tafila Technical University
Tafila, Jordan

Ramiro Marbello[4]
Faculty of Mines Department of Geosciences and Environment, National university of Colombia-Medellin
Medellín, Colombia

***Abstract*—Jordan is suffering a chronicle water resources shortage. Rainfall is the real input for all water resources in the country. Acceptable accuracy of rainfall prediction is of great importance in order to manage water resources and climate change issues. The actual study include the analysis of time series trends of climate change regards to rainfall parameter. Available rainfall data for five stations from central Jordan where obtained from the Ministry of water and irrigation that cover the interval 1938- 2018. Data have been analyzed using Nonlinear Autoregressive Artificial Neural Networks NAR-ANN) based on Levenberg-Marquardt algorithm. The NAR model tested the rainfall data using one input layer, one hidden layer and one output layer with a different combinations of number of neuron in hidden layer and epochs. The best combination was using 25 neurons and 12 epochs. The classification performance or the quality of result is measured by mean square error (MSE). For all the meteorological stations, the MSE values were negligible ranging between $4.32*10^{-4}$ and $1.83*10^{-5}$. The rainfall prediction result show that forecasting rainfall values in the base of calendar year are almost identical with those estimated for seasonal year when dealing with long record of years. The average predicted rainfall values for the coming ten-year in comparison with long-term rainfall average show; strong decline for Dana station, some decrees for Rashadia station, huge increase in Abur station, and relatively limited change between predicted and long-term average for Busira and Muhai Stations.**

***Keywords*—Jordan; rainfall distribution; time series analyses; Levenberg-Marquardt algorithm; climate change**

I. Introduction

Climate change is now becoming a reality rather than hypothesis. Over the last few decades, the atmospheric concentration of carbon dioxide has increase significantly [1]. This increment beside several other factors induced the average temperature of the planet to increase ≈ 0.2 °C per decade in the past 30 years [2]. The effect of global warming on the variability of rainfall is an important issue .Understanding such variability is essential to a reasonable interpretation of a hydrological cycle response to such increase in earth temperature [3] The variability of rainfall is a crucial climate component for society , environment, agriculture, and over all water management plan. Increasing precipitation variability can produce a long series of effects, from reducing the productivity of agriculture to affecting the growth of children [4], [5].

Jordan is considered one of the poorest countries in water resources. Therefore, water resources related issues are always present a source of concern and at the same time a source of interest. Among these issues, we could specify the climate change, groundwater over abstraction, water quality deterioration. Groundwater depletion and precipitation decreasing or precipitation time shifting. As a result, climate change studies are increasingly important [6]. There are historical indications of climate change, or at least there are apparently some areas that used to have Ecosystems that need more water sources than what is available now days [7], [8], [9]. One of the good examples that show these supposed climatic changes is the murals in the Umayyad palaces scattered in the Jordanian deserts. The murals present landscape, fauna and flora that require much water than available resources now days [10], [11]. The Dead Sea is another example of the supposed climate change. The Dead Sea suffers from a continuous decline in its level, and consequently a significant decrease in its area, as well as the disappearance of a number of streams and small rivers that used to discharge to it [12], [13], [14]. The official water resources policy pays great emphasis on climate change, because the country's future and prosperity are related to the abundance of water resources and their adequacy to meet current or expected future requirements [6]. It should be noted that in addition to climate change, the large and irregular increase in population, whether due to natural reproduction or migrations from neighboring countries, complicates the water resources management efforts in Jordan. The rapid growth of population is partially attribute of massive immigration from neighboring countries due to wars and insecurities [15], [16].

Climate change and related issues are of great concern to government agencies, scientific entities and even public masses. By this means, until 2014 the Jordanian ministry of Environment had patronage at least three National Communications on Climate Change. Many scientific articles that deal with climate change, its indicators and impacts have been published especially since 2004 [17], [18].

For common population in Jordan, climate change is of increasing concern. The Dead Sea tragedy presents a chock hup for public awareness regards to climate change. Only in approximately one quarter of hour, a heavy rainfall produced huge flash flood that swept more than 20 people to their death on October 25Th, 2018. On November 10th of the same year another flashflood killed and injured more than 40 people and inundated more than 1000 people in Maan area [19]. This study represent a sample of the scientific response to the massive social worry regards to climate change. Where a good prediction for the pattern and distribution of the future rainfalls will be very helpful for safe management of water resources and anticipate some rainfall related disasters.

In order to forecast the rainfall many methods have been used. Among these methods; simple regression analysis, autoregressive integrated moving average, and exponential smoothing techniques. The accuracy of these methods still in debate until the moment [20], [21], [22]. The development of computing techniques and capabilities enhance the using of certain methods in climate parameters forecasting. The Artificial neural networks is good example of the new techniques where in many studies present a good accuracy [23], [24].

Recently, Artificial intelligent is used in many applications such as power [25], health [26-31], communication [32], text classification [33], [34], texture classification [35],[36], and optimization [37]. Technology and AI applications can be applied in many different sectors and industries to generate maximum production from the operational front. Artificial neural networks are one of the recent trends used in AI applications such as communication [38], wind power prediction [39],[40], text classification [41], civil engineering [42], health [43], image processing [44], climate prediction [45],[46],[47], and power load forecasting [48]. One of the most important features of ANN is its ability to recognize time series data and predict data with high efficiency compared to other methods, especially nonlinear relationships. Predicting the amount of rain that will fall in a given area for a long term is a difficult problem that is still been studied.

There are several theories used in the literature to make long-term prediction, such as linear and nonlinear techniques. Common nonlinear methods are Neural Networks, Support Vector Machine (SVM) and fuzzy logic. Artificial Neural Networks (ANN) has been successfully applied to solve some complex practical problems. ANN can discover and learn relationships between sets of data and link them together. ANN has great capabilities in dealing with nonlinear prediction problems. In addition to that, it has the ability to handle massive numbers of variables. There are several types of them, as mentioned in the following references. After training, it can be used to predict the rainfall.

There has been an increasing interest in the scientific literature regarding accurate prediction in the case of linear and nonlinear systems using artificial neural networks. There are many practical applications in this field. In this scientific paper, the possibility of obtaining annual forecasts of precipitation quantities through neural networks is studied and analyzed. NAR neural networks based on Levenberg-Marquardt algorithm is used here. Prediction results showed a high degree of accuracy of long-term forecasts.

The aims of the actual study are the followings:

- Investigate the availability of rainfall data for a conscious predation study.
- Predict the future rainfall quantity on the base of seasonal year and calendar year in the study area.
- Check the accuracy of ANN techniques for climate forecasting in the study area.
- Observe the climate change pattern related to rainfall parameter in the study area.

II. System Design and Methodology

In this section system design for rainfall amount forecasting is presented. In addition, the used methodology is explained.

A. Study Area and Data Collection

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable. Five rainfall stations were selected in the central region of the Hashemite Kingdom of Jordan (Fig. 1). One of these stations (Mohi) is located in Al-Karak Governorate and the others are in Al-Tafila Governorate. All these stations are located at an altitude of more than one thousand meters above sea level. The areas represented by these stations are the most populated centers in the two governorates.

A large part of the population in these areas practices agriculture, and therefore changes in rainfall levels will have a great impact on the lives of citizens in addition to their impact on the water balance and water resources in general.

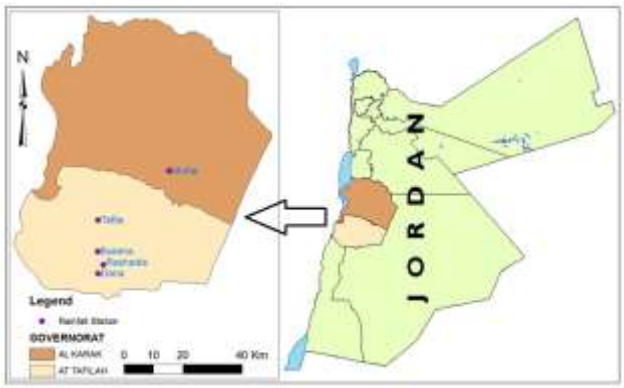


Fig. 1. Study Area Map shows the Spatial Distribution of the Five Rainfall Stationes.

Available rainfall data for the five metrological stations were obtained from the Ministry of Water and Irrigation. The data cover time span extend since the station started working till the 2018 rain season. For some stations the first data back to 1937 where for other the first data backs to 1969 (Table I). The density of data is also a varying issue. Some stations have a good record of data while others have some temporal gapes.

The data include the daily rainfall quantities in mm. The collected data from the five different locations are combined to form one central database. Fig. 2 shows the elevation map of the study area with station ID's.

Table II shows an example of a data from one of the sites (Dana). Fig. 3 shows plot of real rainfall data of the same station from 1945-2010.

B. Data Preprocessing Phase

The pre-process of the data, including data cleaning and missing data treatment. The unnecessary information for rainfall amount model such as Object ID is removed from the database. Then the missing rain in a certain date is replaced by the average value of the rain on the same year. The calendar rainfall values are calculated by summing the daily values from January 1st to December 31st of the calendar year.

TABLE I. TEMPORAL COVERAGE OF RAINFALL DATA FOR THE FIVE STATIONS

Station	Observation period	Comments
Muhi	1968-2017	Sporadic short gaps
Tafila	1938-2018	Data available since October 1937
Busira	1938-2018	Data available since October 1937
Rashadia	1970-2018	Sporadic short gaps
Dana	1946-2011	big gap between 2005-2009

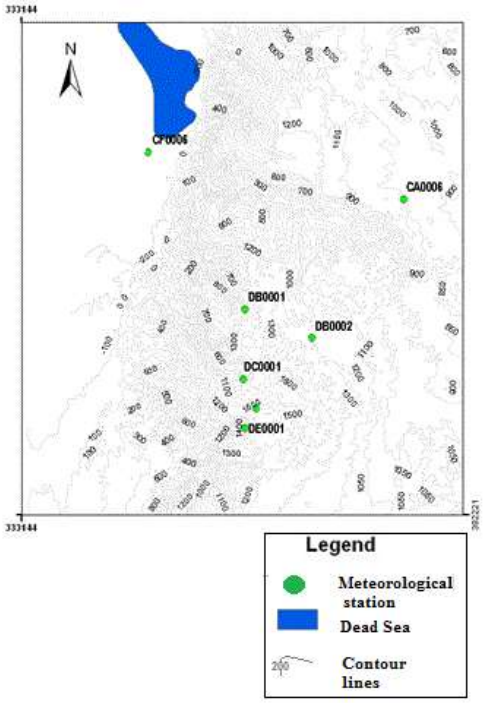


Fig. 2. Elevation Map of the Study Area.

TABLE II. SAMPLE FROM DANA STATION DATABASE

Station ID	Station Name	Date	Rain Amount
DE0001	Dana	17-Nov-1945	20.0
DE0001	Dana	18-Nov-1945	11.3
DE0001	Dana	29-Nov-1945	15.1
DE0001	Dana	1-Dec-1945	2.1
DE0001	Dana	2-Dec-1945	4.0
DE0001	Dana	15-Dec-1945	7.5

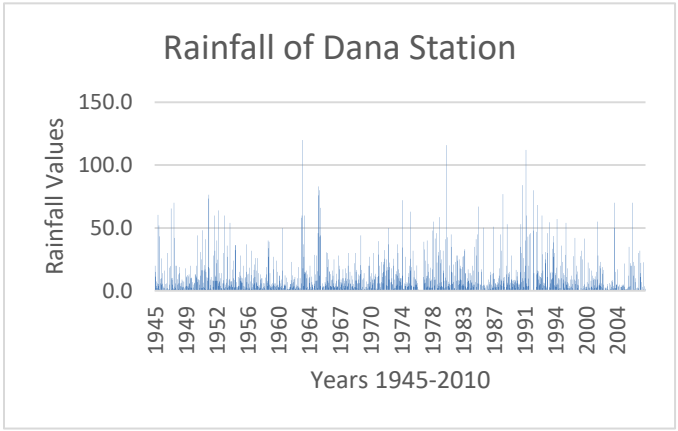


Fig. 3. Plot of Real Rainfall Data of Dana Station from 1945-2010.

The seasonal rainfall values are calculated by summing the daily values from September 1st of a year to August 31st of the followed year. In some studies they suppose that the rain fall season starts in October 1st to September 30 of the next year. Practically this is not issue for the actual study area, where the rainfall from June to end of September is almost zero [16]. Fig. 4 shows the long-term monthly distribution of rainfall in Tafila governorate where practically no considerable rainfall is recorded during June, July, August and September.

C. Nonlinear Autoregressive Artificial Neural Networks Model

The template is designed so that author affiliations are not The NAR network is a feed forward neural network with three layers; input, hidden and output layers. NAR network is used to solve a time series problem. NAR neural network uses historical data of the rainfall in order to do the prediction. The topology of a NAR network is shown in Fig. 5. The number of delays and the number of neurons in the hidden layer are adjustable. These numbers are optimized through trial-and-error testing in order to get accurate model responses.

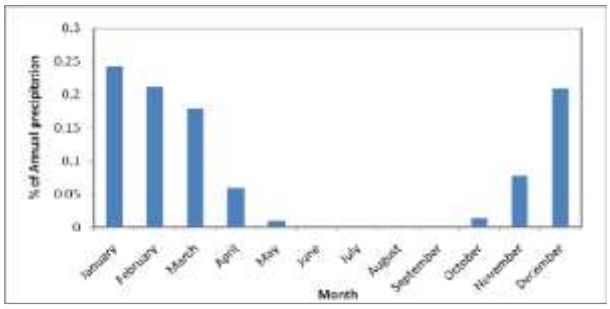


Fig. 4. Long Term Monthly Distribution of Rainfall in Tafila Governorate.

Future values of time series values y(n) are forecasted based on its delays values as shown in (1).

$$y(n) = \sum_{j=1}^{m} w_j * \left[f\left(\sum_{i=1}^{k} w_{ij}\, y(n-i) + w_{0j}\right)\right] + w_0 + \varepsilon \quad (1)$$

Where k is the number of delays, m is the number of neurons in a hidden layers with activation function f, and w_{ij} is the weight of the connection between the input I and the hidden neuron j, w_j is the weight between the hidden neuron j and the output layer. w_{oj} *is* the initial weights (bias) between input layer and j neuron in the hidden layer and and w_o is the initial weight of the output layer.ε is the error of the approximation of the series at a given time. Sigmoid function, a continuous non-linear function, is the most commonly used activation function for neural network design with back propagation training.

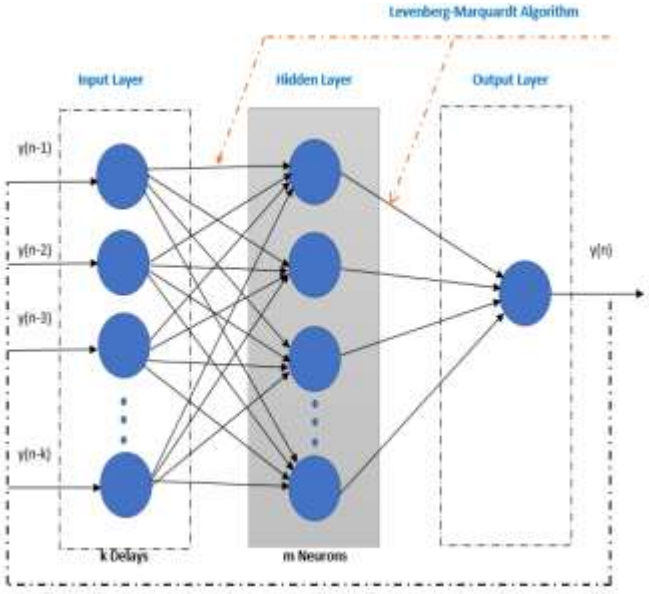


Fig. 5. Nonlinear Autoregressive Artificial Neural Networks ModelTopology.

This formula is used to predict the value of a data series y at time n, y(n), using the p past values of the series. In the training phase, the true output is available and it was used as the input to the network as shown in Fig. 6.

During The training of the neural network aims to approximate the function f by means of the optimization of the network weights and neuron bias. After training, the developed model is used to forecast the rainfall amounts.

The most common learning rule for the NAR network is the Levenberg-Marquardt backpropagation Algorithm. The Levenberg–Marquardt (LM) algorithm is widely used for training the NAR network it has fast convergence speed. The training time of the algorithm is decreased because LM algorithm uses an approximation of the Hessian matrix without direct calculation. The LM equation to update the weights is shown in (2).

$$w_{k+1} = w_k - [J^T J + \delta I]^{-1} J^T e(w_k) \quad (2)$$

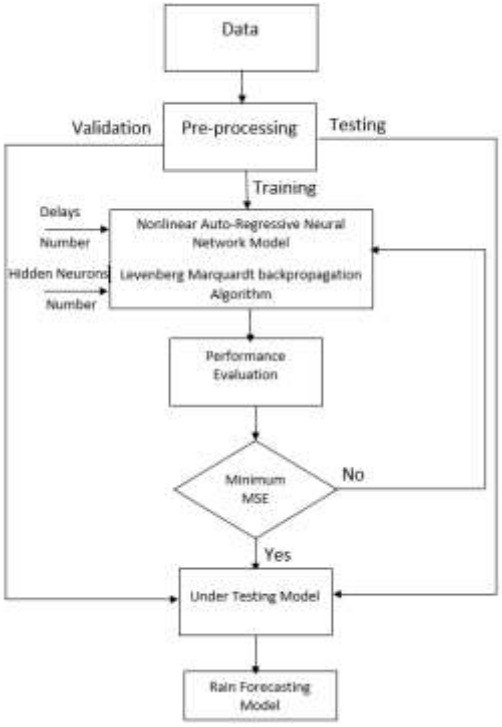


Fig. 6. Developed Rain Forecasting Model using NAR and Levenberg–Marquardt.

Where J is the Jacobian matrix. The Jacobian matrix has the first derivatives of the network error with respect the weights and biases. I is the identity matrix. The variable e is a vector of the network errors in every training sample. The Hessian matrix is $J^T J$ and $J^T e$ is the gradient. The parameter δ is the learning coefficient and is automatically updates based on the error at each iteration. The Levenberg-Marquardt update rule is a blend of both gradient descent and Gauss-Newton iteration. The effectiveness of the model is measured by using mean squared error (MSE) and R Squared in (3) and (4), respectively.

$$MSE = \frac{1}{n}\sum_i (yi - \hat{y})^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum_i (yi - \hat{y})^2}{\sum_i (yi - \bar{y})^2} \quad (4)$$

Where yi is the actual value of the sample i and n is the number of samples, $\hat{y}$ is the forecasted value and $\bar{y}$ is the average value.

III. Result and Discussions

The NAR has been trained using the historical rainfall data over the period 1938-2018 years. Accordingly, the historical data comprising input-target patterns have been divided into three parts: 70% for training, 15% for validation, and 15% as test data. The initial weights in the network are assigned randomly and they were adjusted at each iteration (i.e., epoch) to reduce the error. The procedure continued until the network output met the stopping criteria. The delay of input n = 7. NAR consists of one input layer, one hidden layer, and one output layer. The NARs are based on LM back-propagation training algorithms that were used for long-term time series prediction. MATLAB® software is used to build the models.

The NAR structure with 30 hidden neurons was found to be the most effective. It must be noted that increasing the number of neurons in the hidden layer makes a system more complex and decreasing the amount of neurons in the hidden layer will lower the computing power and generalization of the ANN. The results shows clearer trends using seasonal year data sets in comparison with calendar year data sets as shown in Fig. 7 to Fig. 10 for two locations Dana and Muhai.

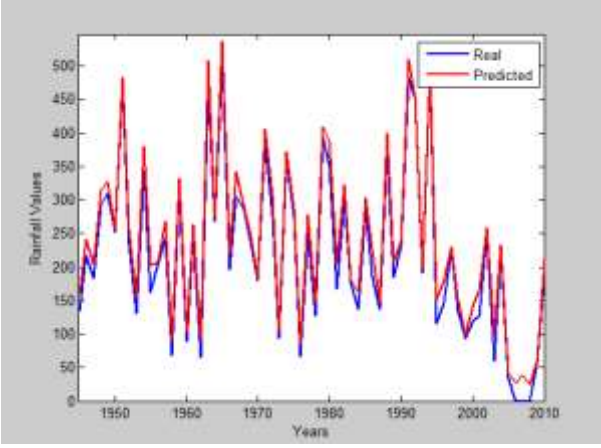


Fig. 7. Annual Rainfall Values Forecasting Dana Location.

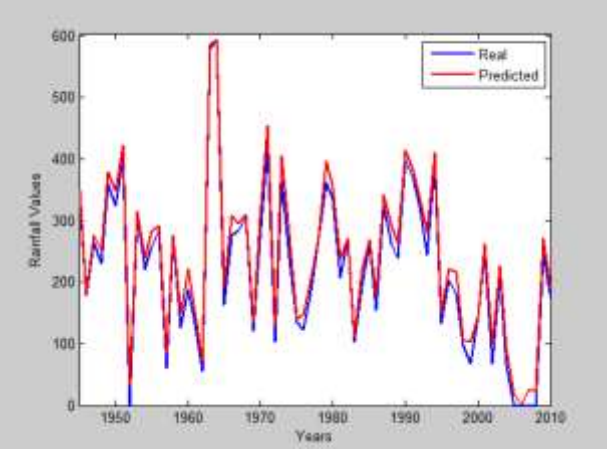


Fig. 8. Seasonal Rainfall Values Forecasting Dana Location.

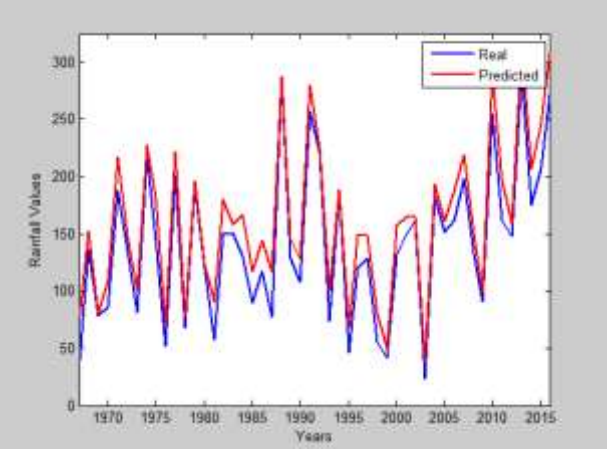


Fig. 9. Annual Rainfall Values Forecasting Muhai Location.

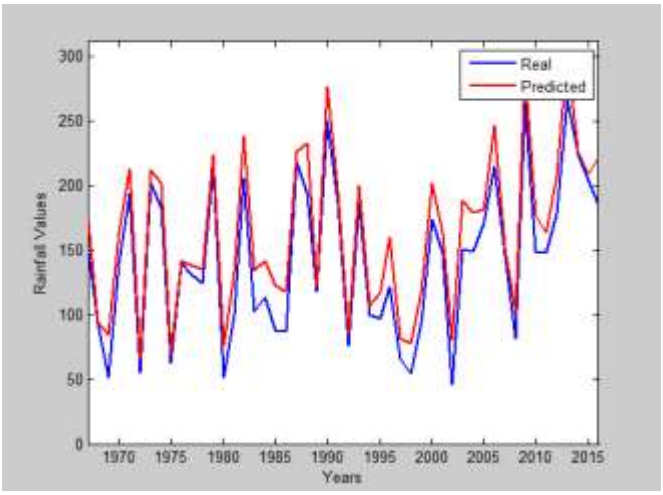


Fig. 10. Seasonal Rainfall Values Forecasting Muhi Location.

Both of prediction sets has very good accuracy where the value of mean square error (MSE) between the actual response and the target for test data is less than $5*10^{-4}$ in all cases (Table III). The MSE error obtained can be considered negligible for nonlinear systems and it is less than what achieve in some other studies which enhance the credibility of rainfall prediction of the actual study. The annual prediction results shows that using 25 neurons in hidden layer, 5 delay and 12 epoch produces predictions that perfectly match the input data (Table IV). This table present the prediction results of NAR NN with n=25 neurons in hidden layer and changing the number of delay.

Table IV show that R=1 when delay equal 5, so NAR performs best when delay is set 5. This means 5 earlier years' prediction is meaningful to future prediction. When the delay is larger than 5, It will cause over fit issue which make the trained network less adaptable.

TABLE III. MEAN SQUARE ERROR (MSE) BETWEEN THE ACTUAL RESPONSE AND THE TARGET FOR TEST DATA

MSE	Calendar Year	Seasonal Year
CA0006	$4.32*10^{-4}$	$4.67*10^{-4}$
DB0002	$1.83*10^{-5}$	$4.92*10^{-6}$
DC0001	$1.42*10^{-4}$	$1.09*10^{-4}$
DC0002	$3.34*10^{-5}$	$2.12*10^{-4}$
DE0001	$1.34*10^{-4}$	$2.14*10^{-4}$

TABLE IV. NUMBER OF DELAYS AND EPOCH AND R OF DANA LOCATION

Delay	Epoch	R
1	22	0.32
3	15	0.93
5	12	1
8	43	0.8
10	91	0.63
12	31	0.78
15	75	0.48

Prediction results of NAR NN with delay=5 and changing the Number of neuron in hidden layer of Dana location is shown in Table V. In this table it is obvious that NAR performs best with number of neuron in hidden layer is set 25. This means 25 hidden is meaningful to future prediction. When number of neurons is larger, the NAR become worse.

Some models with different delays and hidden neurons combinations are shown in Table VI. This table shows that the model with 5 delay and 25 hidden neurons offers the best prediction accuracy where the value of Training R, Testing R and Validation R equal 0.98 or more.

The prediction in the time spam where the data is available is to validate the method and to check the model accuracy, which is an important issue. However, expanding the adapted method to cover certain time in the future is the real purpose of prediction. In this study, the future forecasting cover ten years beyond the last available rainfall data.

TABLE V. NUMBER OF NEURON IN HIDDEN LAYER AND EPOCH AND R OF DANA LOCATION

Number of neuron in hidden layer	Epoch	R
3	55	0.25
5	40	0.33
9	81	0.45
15	14	0.64
20	22	0.78
25	12	0.98
30	27	0.89

TABLE VI. RESULT OF SOME MODEL WITH DIFFERENT DELAYS AND HIDDEN NEURONS COMBINATIONS

Model Structure	Training R	Testing R	Validation R
3 delay, 5 hidden neurons	0.34	0.22	0.1
8 delay, 15 hidden neurons	0.54	0.31	0.2
2 delay, 20 hidden neurons	0.78	0.62	0.44
5 delay, 25 hidden neurons	1	0.98	0.99
10 delay, 30 hidden neurons	0.88	0.62	0.48

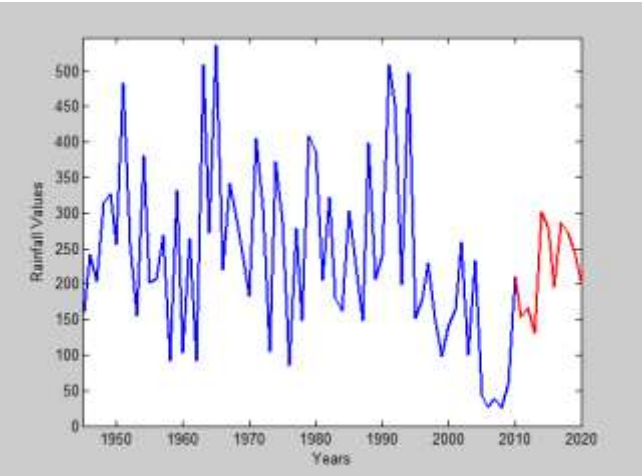


Fig. 11. Annual Rainfall Values Forecasting Dana Location.

Fig. 11 to Fig. 14 show the plot results of predicting for 10 years of two locations for the best-achieved model. The red colored line represents the predicted rainfall values for 10 years.

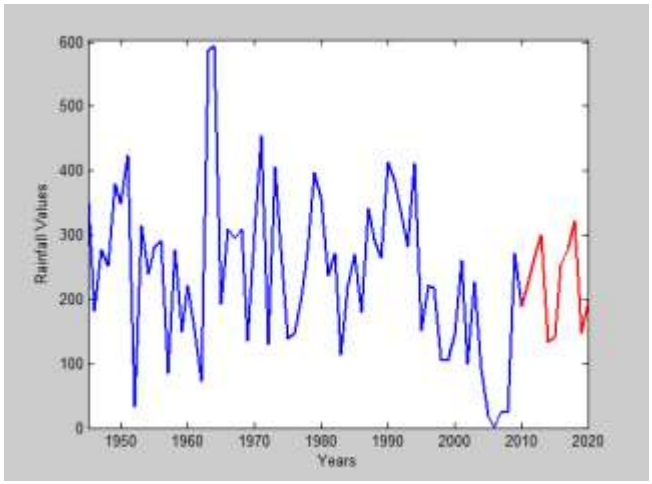


Fig. 12. Seasonal Rainfall Values Forecasting Dana Location.

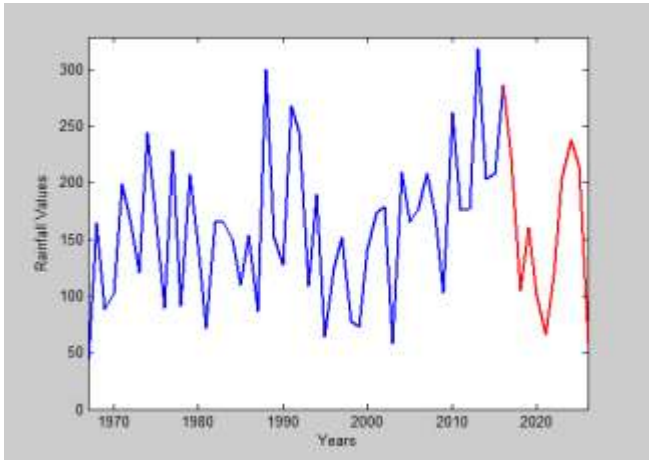


Fig. 13. Annual Rainfall Values Forecasting Muhi Location.

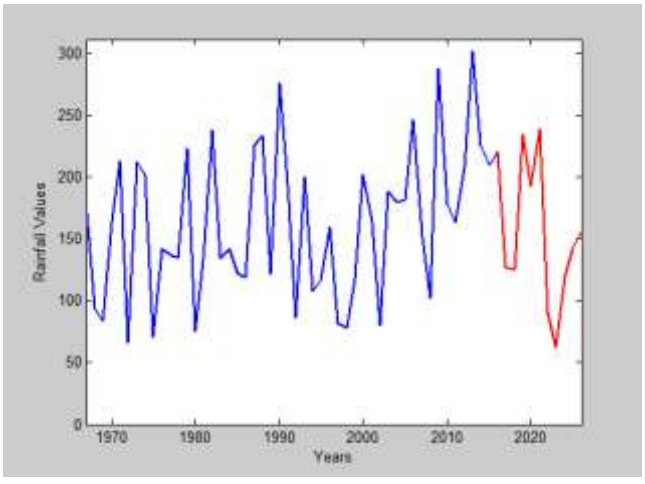


Fig. 14. Seasonal Rainfall Values Forecasting Muhi Location.

The data presented in the Fig. 11 to Fig. 14 are summarize in Table VII. This table list long term recorded rainfall average and the rainfall average of the ten predicted years for each location. As the long-term average is almost the same in case of seasonal or calendar year base, the table prepared using the calendar year rainfall data. The comparison of average ten years predicted rainfall with the long term average revile that for Tafila station the predicted value is much higher than long term average (371 vs 286). For Muhi and Busira stations, the predicted values are quiet higher than the long-term average (149 vs 141 and 238 vs 234 respectively). For Rashadia station, the predicted rainfall is less than the long-term average (203 vs 218). While for Dana station the ten years average prediction is about 20 % less than long-term rainfall average (286 vs 236). The difference between the long-term average rainfall and the average of predicted rainfall for the coming ten years seem to have a spatial pattern.

TABLE VII. LONG TERM RAINFALL AVERAGE AND AVERAGE OF 10 YEARS PREDICTED RAINFALL VALUES FOR THE FIVE STATIONS (MM / YEAR)

	Station				
	Muhi	Tafila	Busira	Rashadia	Dana
Long term average	142	286	234	218	286
Predicted 10 years average	149	371	238	203	230

In Tafila station, which locate in the center of study area the prediction, expect high increase (Fig. 15). In the next stations toward north and south (Muhi and Busira) the prediction is some increase. Moving more toward south to Rashadia station, the expectation is decrease. While in Dana station which is the most southern station, the expectation is high decrease.

This study showed that there is no clear trend for future expectations, as it is evident from the above that some stations are expected to receive more precipitation, some less and some of them are almost stable. This result is consistent with the findings of [49] and [17], that there is no specific trend in their study areas. With regard to spatial related changes, many studies agree with this research, where they expected different changes in different places, but many of these studies used short time coverage data, and since precipitation rates in Jordan in general are highly variable, changing the length of the record may lead to different results [18].

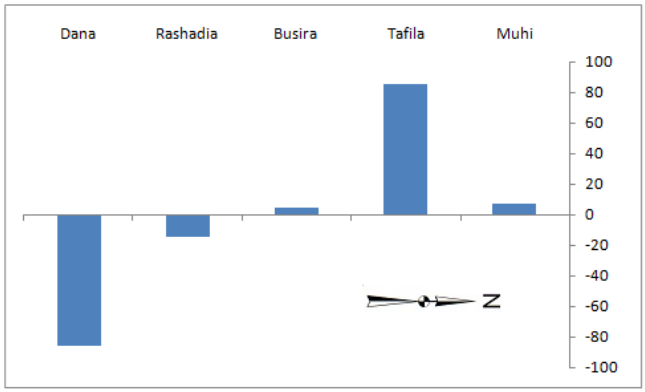


Fig. 15. Average 10 Year Predicted Rainfall – Long Term Average for the Five Stations in the Study Area (mm).

IV. CONCLUSIONS

The results of the study show that the rainfall forecast efficiency is nearly identical for both seasonal and calendar years. The ANN technique proves to be highly efficient and accurate, and the MSE values were less than $5 * 10^{-4}$ in all cases. The best accuracy was achieved with 25 hidden neurons and 5 delays. The results obtained showed that the future projection of precipitation is not uniform in the whole region. The expected rainfall amounts showed relative stability in two stations, a low decrease in one station, a significant decrease in one station, and a high increase in one station. According to the long-term mean, the projected shift in rainfall shows a spatial pattern, as there is a very large increase in the central station (Tafila) and the highest projected decline in the southern station (Rashadiya).

REFERENCES

[1] Keeling, R.F.; Piper, S.C.; Bollenbacher, A.F.; Walker, J.S. Atmospheric CO2 records from sites in the SIO air sampling network. In Trends: A Compendium of Data on Global Change; Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy: Oak Ridge, TN, USA, 2009.

[2] Hansen J, Sato M, Ruedy R, Lo K, Lea DW, Medina-Elizade M., Global temperature change. PNAS, vol. 103, no.39, 2006.

[3] Pendergrass AG, Knutti R, Lehner F, Deser C, Sanderson BM. Precipitation variability increases in a warmer climate. Sci. Rep. ; vol. 7, 2017.

[4] Rowhani P, Lobell DB, Linderman M, Ramankutty NClimate variability and crop production in Tanzania. Agric. For. Meteorol, 2010.

[5] Shively GEInfrastructure mitigates the sensitivity of child growth to local agriculture and rainfall in Nepal and Uganda. Proc. Natl., 2017.

[6] Ministry Of Environment –MOE-,The National Climate Change Policy of the Hashemite Kingdom of Jordan 2013-2020: Sector Strategic Guidance Framework., 2013.

[7] Shehadeh, N., The Climate in Jordan in the Past and Present. In. Hadidi, A. (Editor). Studies in the History and Archaeology of Jordan, II. Published by the Department of Antiquities. Amman, Jordan., 1985.

[8] Mountfort, G., Portrait of A Desert, The Story of an Expedition to Jordan. Collins. St James.s Palace, London, 1965.

[9] Nelson, J.B., Azraq Desert Oasis. Allen Lane. London, Fakenham and Reading., 1973.

[10] Alawneh, F., Balaawi, F., Hadad, N., and Shawabkeh, Y., "Analytical identification and conservation issues of painted plaster from Qaser Amra in Jordan," International Journal of Conservation Science. vol. 2, no. 4, 2011.

[11] Nelson, J.B., Azraq . A Case Study. In. Hadidi, A, " Studies in the History and Archaeology of Jordan, II" Published by the Department of Antiquities. Amman, Jordan., 1985.

[12] Alpert, A., Shafir, H., and Issahary, D., "Recent Changes in the Climate At the Dead Sea – a Preliminary Study," Climate Change, vol. 37, no. 3, 1997.

[13] Abu Ghazleha,S., Kempea,S., Hartmannb, J., and Jansenb, N.," Rapidly Shrinking Dead Sea Urgently Needs Infusion of 0.9 km3/a from Planned Red-Sea Channel: Implication for Renewable Energy and Sustainable Development. Jordan," Journal of Mechanical and Industrial Engineering, vol.4, no. 1, 2010.

[14] Al Eisawi, D .M.H., "Water scarcity in relation to food security and sustainable use of biodiversity in Jordan," Food security under water scarcity in the Middle East:Problems and solutions, vol 2, pp. 239 -248 , 2005.

[15] Ramírez, O.A ., Ward, F. A., Al-Tabini, R. and Phillips, R., Efficient water conservation inagriculture for growing urban water demands in Jordan, Water Policy Vol 13 No 1 pp 102–124. (2011):

[16] Sharadqah, S., "Climate Change Trends in Tafila Governorate (Central West Jordan) in the Period 1938- 2006," Journal of Natural Sciences Research. vol. 4, no.10, p 23-35, 2014.

[17] Dahamsheh, A. and Aksoy, H. “Structural characteristics of annual precipitation data in Jordan,” Theor. Appl. Climatol, vol. 88, pp. 201–212, 2007.

[18] Matouq, M.,El-Hasan, T., and Al Bilbisi. H, “The climate change implication on Jordan: Acase study using GIS and Artificial Neural Networks for weather forecasting,” Journal of Taibah University for Science, vol 7, no. 2, pp. 44-55, 2013.

[19] Hana Namrouqa, “ Floods 'only beginning' of severe climate change impacts on Jordan,” The Jordan Times. Last updated at vol. 11, 2018.

[20] Haviluddin and R. Alfred, “ Forecasting Network Activities Using ARIMA Method,” Journal of Advances in Computer Networks, vol. 2, no.3, pp. 173-179. 2014.

[21] Shrivastava, G., S. Karmakar, and M.K. Kowar, “Application of Artificial Neural Networks in Weather Forecasting: A Comprehensive Literature Review,” International Journal of Computer Applications, vol. 51, no.18, pp. 17-29, 2012.

[22] Farajzadeh, J., A.F. Fard, and S. Lotfi, “Modeling of monthly rainfall and runoff of Urmia lake basin using “feed-forward neural network” and “time series analysis” model, “ Water Resources and Industry, vol. 7, no. 8, pp. 38–48, 2014.

[23] Abhishek, K., et al, “A Rainfall Prediction Model using Artificial Neural Network,”,IEEE Control and System Graduate Research Colloquium, 2012.

[24] Charaniya, N.A. and S.V, “Dudul, Design of Neural Network Models for Daily Rainfall Prediction,” International Journal of Computer Applications, vol. 61, no.14, pp. 23-26, 2013.

[25] Mansour, A.M., Abdallah, J., Obeidat, M.A.,” An efficient intelligent power detection method for photovoltaic system,” International Journal of Circuits, Systems and Signal Processing, vol. 14, pp. 686–699, 2020.

[26] Ayman M. Mansour, Murad M. Alaqtash, Mohammad Obeidat “Intelligent Classifiers of EEG Signals for Epilepsy Detection,” WSEAS Transactions on Signal Processing, vol. 15, 2019.

[27] Murad Alaqtash, Ayman M Mansour, Mohammad Obeidat, “Fuzzy Assessment Model for Functional Impairments in Human Locomotion”. IOSR-JECE, vol. 14, no. 1, Jan-Feb 2019.

[28] Ayman M. Mansour, “Intelligent E-Health System for Patient and Elderly People Monitoring Using Multi Agents System,” Jordan Journal of Electrical Engineering, vol. 4, no. 1, 2018.

[29] Ayman M. Mansour, “Decision Tree-Based Expert System for Adverse Drug Reaction Detection using Fuzzy Logic and Genetic Algorithm,” International Journal of Advanced Computer Research (IJACR), vol. 8, no. 36, 2018.

[30] Mohammad A. Obeidat and Ayman M. Mansour, “EEG Based Epilepsy Diagnosis System using Reconstruction Phase Space and Naïve Bayes Classifier,” WSEAS Transactions on Circuits and Systems, vol. 17, 2018.

[31] Mansour, A.M., Obaidat, M.A. and Hawashin, B. Elderly people health monitoring system using fuzzy rule based approach. International Journal of Advanced Computer Research, vol. 4, no. 4, p.904. 2014.

[32] Ayman M. Mansour, “GSM based Vehicle-to-Vehicle Communication using Multi-Agent Intelligent System,” WSEAS Transactions on Electronics, vol. 10, 2019.

[33] B. Hawashin et al., "Efficient Texture Classification Using Independent Component Analysis," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 544-547, 2019.

[34] Bilal Hawashin, Ayman M. Mansour and Shadi Aljawarneh, “An Efficient Feature Selection Method for Arabic Text Classification,” International Journal of Computer Applications (IJCA), vol. 83, no. 17, pp. 1-6, December 2013.

[35] Ayman M. Mansour, “Texture Classification using Naïve Bayes Classifier,” International Journal of Computer Science and Network Security (IJCSNS), vol. 18, no. 1, January 2018

[36] D.A. Al Nadi and Ayman Mansour, "Independent Component Analysis (ICA) for texture classification", 5th International Multi-Conference on Signals and Devices, IEEE SSD, 2008.

[37] Jafar.Abu Khait, Ayman M Mansour and Mohammad Obeidat, “Classification based on Gaussian-kernel Support Vector Machine with Adaptive Fuzzy Inference System,” PRZEGLĄD ELEKTROTECHNICZNY., vol 5, pp 16-24, 2018.

[38] Ayman M Mansour, “ Cooperative Multi-Agent Vehicle-to-Vehicle Wireless Network in a Noisy Environment,” International Journal of Circuits, Systems and Signal Processing, vol. 15, 2021.

[39] Yu, M., Zhang, Z., Li, X., Yu, J., Gao, J., Liu, Z., Yu, R. “Superposition Graph Neural Network for offshore wind power prediction. Future Generation Computer Systems,” vol. 113, pp 145-157. 2020.

[40] Wang, S., Zhao, X., Wang, H., & Li, M., “ Small-world neural network and its performance for wind power forecasting,” CSEE Journal of Power and Energy Systems, vol.6, no. 2, pp. 362-373, 2019.

[41] Murray, Fraser. "Text Classification using Artificial Neural Networks." pp 1-70, 2015.

[42] C. Xu, B. Gordan, M. Koopialipoor, D. J. Armaghani, M. M. Tahir and X. Zhang, "Improving Performance of Retaining Walls Under Dynamic Conditions Developing an Optimized ANN Based on Ant Colony Optimization Technique," in IEEE Access, vol. 7, pp. 94692-94700, 2019.

[43] B. Xiong et al., "Intelligent Prediction of Human Lower Extremity Joint Moment: An Artificial Neural Network Approach," in IEEE Access, vol. 7, pp. 29973-29980, 2019.

[44] X. Zhou et al., "A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks," in IEEE Access, vol. 8, pp. 90931-90956, 2020.

[45] Antonić, O., Križan, J., Marki, A., & Bukovec, D, “patio-temporal interpolation of climatic variables over egion of complex terrain using neural networks,” Ecological Modelling, vol. 138, pp. 255–263, 2001.

[46] Boulanger, J.-P., Martinez, F., & Segura, E. C., “Projection of future climate change conditions using IPCC simulations, neural networks and Bayesian statistics Part I: Temperature mean state and seasonal cycle in South America, “ Climate Dynamics, vol. 27, pp. 233–259, 2006.

[47] Boulanger, J.-P., Martinez, F., & Segura, E. C.,” Projection of future climate change conditions using IPCC simulations, neural networks and Bayesian statistics. Part II: Precipitation mean state and seasonal cycle in South America,” Climate Dynamics, vo. 28, pp. 255–271, 2007.

[48] H. Dong, Y. Gao, X. Meng and Y. Fang, "A Multifactorial Short-Term Load Forecasting Model Combined With Periodic and Non-Periodic Features - A Case Study of Qingdao, China," in IEEE Access, vol. 8, pp. 67416-67425, 2020.

[49] Al Qatarneh, G.N., Al Smadi, B., Al-Zboon, K., “ a Impact of climate change on water resources in Jordan: a case study of Azraq basin,” Appl Water Sci, vol. 8, no. 50, 2018.

Fungal Blast Disease Detection in Rice Seed using Machine Learning

Raj Kumar[1], Gulsher Baloch[2], Pankaj[3], Abdul Baseer Buriro[4], Junaid Bhatti[5]
Department of Electrical Engineering, Sukkur IBA University, Sukkur, Pakistan

***Abstract*—The economy of Pakistan mainly relies upon agriculture alongside other vital industries. Fungal blast is one of the significant plant diseases found in rice crops, leading to reduction of agricultural products and hindrance in the country's economic development. Plant disease detection is an initial step towards improving the yield and quality of agricultural products. Manual Analyzation of plant health is tiresome, time taking and costly. Machine learning offers an alternate inspection method providing benefits of automated inspection, ease of availability, and cost reduction. The visual patterns on the rice plants are processed using the machine learning classifiers such as support vector machine (SVM), logistic regression, decision tree, Naïve Bayes, random forest, linear discriminant analysis (LDA), principal component analysis (PCA), and based on classification results plants are recognized as healthy or unhealthy. For this process, a dataset containing 1000 images of rice seed crop is collected from different fields of Kashmore, and whole analysis of image acquisition, pre-processing, and feature extraction is done on the rice seed only. The dataset is annotated with healthy and unhealthy samples with the help of a plant disease expert. The algorithms used for processing data are evaluated in terms of F1-score and testing accuracy. This paper contains results from traditional classifiers, and alongside these classifiers, transfer learning has been used to compare the results. Finally, a comparative analysis is done between the results of traditional classifiers and deep learning networks.**

Keywords*—Fungal blast; machine learning; support vector machine (SVM); logistic regression; decision tree; Naïve Bayes; random forest; linear discriminant analysis (LDA); principal component analysis (PCA); image acquisition; pre-processing; feature extraction; F1-Score; convolutional classifier; deep learning

I. Introduction

Rice is one of the major agricultural crops in Pakistan, which has a great influence on the country's economy. It is subject to different diseases in its leaves, root, and seed, which may reduce its yield and lead to a reduction in agricultural products [1]. Farmers do not have a specific idea regarding pesticides as per diseases on rice crops [2]. Hence, the rice seed health monitoring with the help of image processing and machine learning algorithms plays an important role in increasing the yield and production of rice [3]. Different related work has been done using machine learning algorithms on rice as well as other crops, which is discussed in Section II. The uniqueness of this research is a dataset of rice seed which is mentioned in Section V. The image processing helps to visualize the plant's images clearly while removing the extra background and extracting the infected region of the plant with the help of feature extraction and segmentation [4]. All the image processing and classification techniques that have been used are mentioned in the proposed workflow in Section III. Machine learning helps to analyze the plant's health based on the extracted features or cropped images of the dataset [4] [5]. With the help of this process, a disease can be detected in rice crops, and based on that disease, farmers can use the specific pesticide, which will lead to a reduction in cost and time [5]. The current methods for rice disease detection in Pakistan involve the experience of farmers in detecting rice disease, which is not very reliable. Further, the inspection by the disease detection expert is too costly, and local farmers are unable to afford it. This, in turn, affects the production and yield of the rice crops. With the recent advancement in machine learning, this paper proposes the vision-based approach to detect rice plant disease. One of the critical requirements of any machine learning problem solution is data generation and collection. Further, for the machine learning technology to be implemented in real-time requires the handling of different image vision problems such as occlusion detection, background/foreground detection, suitable feature selection, and extraction from the rice crop images to complete the required disease detection task. To imitate the real-time solution implementation, the image data of rice plant from a number of different rice fields in Kashmore, city of Pakistan, has been collected, and with the help of a disease detection expert, the data set has been labeled. Further, recent and state-of-the-art machine learning algorithms are implemented and tested on the dataset for rice disease detection, and the results are compared in terms of F-score and accuracy. All the proposed work which successfully have been implemented is mentioned in section IV, which has multiple results. A final conclusion has been made over different classification results, which is mentioned in Section VI. This paper helps summarize the recent and state-of-the-art algorithms for rice disease detection and also helps the authors to cater upon problems for the implementation of the algorithms in real-time rice disease detection.

II. Literature Review

Kawcher Ahmed et al. [7] implemented a machine learning algorithm for the detection of three common rice leaf diseases which are leaf smut, bacterial leaf blight, and brown spot diseases. The dataset used was already refined and collected from an online website [8]. For classification purposes, KNN (K-Nearest Neighbor), Decision Tree, Naïve Bayes, and Logistic Regression [8] [9] are used. It is concluded that the decision tree algorithm after 10-fold cross-validation has better performance with an accuracy of 97% applied on the test dataset. Neha G. Kurale et al. [9] analyzed

leaf diseases in plants generally using the texture features and neural network. They summarized that for the plant's leaf disease detection, support vector machine (SVM), KNN (K-Nearest Neighbor), and Neural Networks techniques [9] have the most appropriate and effective results. Anjna et al. [10] have worked on capsicum disease symptoms, and she has used k-means clustering, BPNN classifier, neural network classifier, thresholding-based segmentation, minimum distance criterion, and SVM [10]. The authors have extracted GLCM features on which they have classified the capsicum of diseases. The SVM and KNN classifiers have 100% accuracy being the highest [10]. It is concluded that neural network classifier gives better results as compared to others in a short time with texture, shape, and co-efficient features [11] [12]. Naga Swetha R. et al. [13] analyzed and detected four different diseases in rice plants which are the bacterial blight of rice, rice blast, and false smut. The dataset of total 115 rice disease images have been collected by themselves and some have been collected from the internet [13]. Only two classifiers, support vector machine (SVM) and KNN (K-Nearest Neighbor) are used based on shape and color features [13]. A mobile application for the automatic diagnosis of diseases in rice plants has also been developed [13]. Muhammad Kashif et al. [14] analyzed the different feature techniques regarding plant disease detection generally. The authors used the texture, Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), Binary Robust Invariant Scalable Keypoints (BRISK), Binary Robust Independent Elementary Features (BRIEF), and Fast Retina Keypoints (FREAK) features for plant disease detection [14]. They concluded that dense SIFT features give the best results with an accuracy of 98.36% [14] [15]. Harshadkumar B. Parjapati et al. [16] analyzed and implemented a machine learning algorithm for the three different leaf diseases of rice plants which are bacterial leaf blight, brown spot, and leaf smut. They collected datasets from the rice fields [16]. They have applied three different techniques of segmentation and for the accurate features, they used K-means clustering segmentation [16] [17]. For the classification, they used an SVM classifier based on color, shape, and texture features [17]. They got an accuracy of 93.33% for the training dataset and 73.33% for the testing dataset. They have also applied k-fold cross-validation and got an accuracy of 83.80% for 5-folds and 88.57% for 10-folds [17]. Efetkhar Hossain et al. [18] used only KNN (K-Nearest Neighbor) classifier based on texture features for the detection of plant diseases. They used the dataset of 237 plant leaf images that were already refined and have been collected from two different database websites [18]. They proposed that the KNN classifier can classify the diseases like Alternaria alternate, anthracnose, bacterial blight, leaf spot, and canker of various plant species [18]. They concluded that the proposed KNN classifier with texture features could detect diseases with 97.76% accuracy [18]. Budiarianto Suryo Kusumo et al. [19] proposed a machine learning algorithm for disease detection in the Corn crop. The dataset used was already refined and has been collected from the PlantVillage dataset website [19]. They used several image processing techniques for feature extraction such as SIFR, SURF, BRIEF, and HOG [19][20][26]. For classification purposes, they used SVM, decision tree, random forest, and Naïve Bayes algorithms [20]. Finally, it is concluded that the color features are most important for disease detection in the corn crop. Sandeep Kumar et al. [21] used support vector regression (SVR) with different classification based on shape, color, texture, and cosine features of plant species for plant disease detection. The authors used a limited plant leaf dataset that has been collected by themselves. They proposed three different computer vision techniques for plant disease detection which are feature discovery, feature explanation, and image depiction [5] [4] [13]. The proposed approach uses SIFT and SURF features and the clustering is done by F-Dbscan [5]. Sachin D. Khirade et al. [1] discussed the different techniques and processes for plant health monitoring and disease detection. The dataset they used, is captured by themselves [1]. They proposed the image processing techniques such as image pre-processing and image segmentation are the most useful for plant disease detection [4] [6] [7]. They used different feature extraction techniques for the extraction of texture, shape, and color features. For classification purposes, they used ANN (Artificial Neural Network) such as self-organizing feature map, back propagation algorithm, and SVM [12]. Pushkara Sharma et al. [19] conducted a study in India on various plant leaves to detect the diseases using pre-processing techniques and segmentation to get the useful part of the leaf. After preprocessing and segmentation, they used Logistic regression, KNN, SVM, and CNN classifiers [19]. The highest accuracy that he got was 98.0% from the CNN model. The authors proposed that through segmentation, the diseased portion of the input image can be detected [21]. For the feature extraction, different feature extraction techniques and different classifiers are used. Arsa, D. M. S et al. [22] has used VGG-16 pre-trained model in Batik based on random forest. They have used precision, recall, F-score, and accuracy to evaluate their proposed method performance [22]. Ufaq Khan et al. [25] divided plant disease detection techniques into two phases; the first is segmentation, and the other is classification. In this paper, the author generally described the techniques for plant disease detection, so they did not use any dataset [25].

After reviewing all the mentioned studies, the proposed work is novel because in the above studies, mostly plant dataset used consists of less than 300 images from one field, and mostly dataset has been collected from the internet, which was already refined and did not necessarily reflect the real field scenario. But in this case, the unique dataset of 1000 healthy and unhealthy rice seed images have been captured from different rice fields. Another uniqueness from the above studies is that most have extracted limited image features while in this case, three different types of features of an image, such as texture, SURF, and BRISK features have been extracted. Moreover, for the testing and training results in the above studies, limited classifiers have been used, such as SVM and decision tree, while in this case, six different classification algorithms such as SVM, LDA, decision tree, logistic regression, Naïve Bayes, and random forest have been used. For the most accurate results, the dataset has been used with different image sizes such as 128x128, 256x256, 512x512, and 1024x1024. PCA and k-fold cross-validation have been applied to every classifier for better accuracy

performances, and finally, the comparatively better results are with SVM and random forest classifiers.

III. Proposed Methodology

In this section, a complete methodology for fungal blast disease detection has been proposed in a block diagram, shown in Fig. 1. Every step is performed for the best accuracy results. In image acquisition, a unique dataset has been collected, and different image processing techniques are applied, such as image cropping, color enhancement, and image resizing for a better understanding of the dataset. Further, feature extraction techniques are used, such as BRISK, SURF, and texture features, to remove the extra background and to get the infected region of dataset. The extracted features are used for the classification purposes while taking 80% of the training dataset and 20% of the testing dataset. Different classifiers such SVM, LDA, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes and PCA classifiers with 10-fold cross validation are used for a comparative analysis based on F1-score and testing accuracy. A descriptive analysis is given as under:

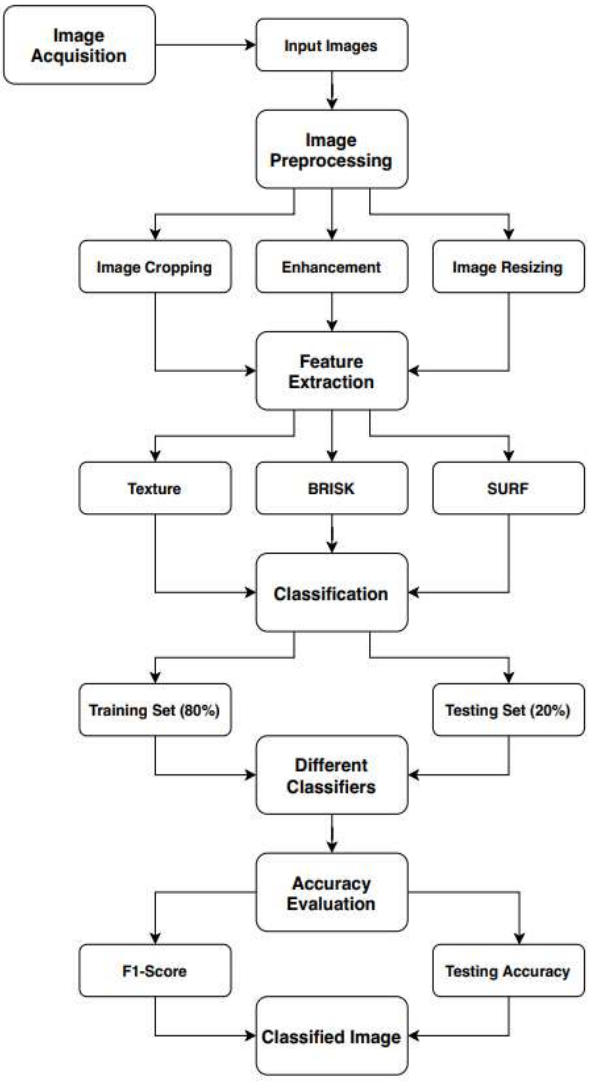


Fig. 1. Proposed Work Flow Chart for Traditional Classification.

A. Image Acquisition

A unique dataset of healthy and unhealthy rice crop has been captured through an android camera from the different fields of Kashmore. The dataset has been captured from September 5th to 7th, 2020 and the age of the crop at that time was 50 to 60 days. The captured images were in RGB (Red, Green, and Blue) form. The whole dataset consists of both healthy and unhealthy crops of 1300 different data samples of rice seed plants annotated with the help of a plant disease expert.

1) Dataset description: Initially, a total of 1500 images of healthy and unhealthy rice crops have been captured from the field. Due to huge distortion in the background and extra parts, images that were not helpful have been removed, and finally, the 1000 healthy and unhealthy images are left in the dataset. The dataset is uploaded on "Kaggle" website, which is now open to use for everyone. The sample images of the healthy and unhealthy dataset are shown in Fig. 2.

Fig. 2. Sample Images from the Dataset of Rice Seed Plant uploaded on Kaggle: (a) and (b) are Healthy Plants of Rice Crop While (c) and (d) are unhealhty Crop because it has Brownish Spots on Seeds.

B. Image Pre-Processing

different pre-processing techniques have been used to prepare data for machine learning classification and evaluation, such as image cropping, image resizing, and image enhancement [20] [21] [22].

1) Image cropping: Image cropping has been performed manually for every image to remove the extra part from the images [17] [18] [20].

2) Image resizing: Image resizing has been done to take all the datasets of equal size, which will help in feature extraction to get the balanced features [6] [7] [13]. For the comparison of better results, all the image sizes have been taken, such as 128 x 128, 256 x 256, 512 x 512, and 1024 x

1024. Better results have been achieved for the image size of 256 x 256 in every classifier. The comparison histograms are shown in Fig. 6.

3) Image enhancement: Image enhancement has been performed for the whole dataset to increase the contrast of images. The RGB dataset has been converted into grayscale for better performances [6] [7] [13].

C. Feature Extraction

Feature extraction is a key step to analyze the image deeply with the help of features. It helps to get useful information from the image [1]. Multiple features of the rice plant dataset have been extracted, such as Gray level co-occurrence matrix (GLCM) or texture features [1] [3], brisk and surf features [14] [3], shown in Table I. This table shows the feature types and their name that have been extracted from the dataset of rice crops. A total of three feature types have been extracted, and normalization is applied for all the features before classification.

1) Texture features: Texture features define the distribution of color, roughness, and hardness in an image. It helps mainly for the detection of infected areas in the image of rice crop [5]. Texture-based features are contrast, correlation, energy, entropy, and homogeneity [14]. Contrast is the intensity measurement between a pixel and its neighbor in an image. Correlation defines that how correlated a pixel is with its neighboring pixel in the entire image. Energy is the measurement of uniformity which means how much homogeneous an image is, the large the energy. Entropy is the measurement of image intensity or disorder. Homogeneity defines the similarity of pixels in an image [13]. Equations for all the texture features or gray level co-occurrence matrix of these features are shown in Table II.

2) Speeded-Up Robust Features (SURF): The SURF algorithm is related to the Scale Invariant Feature Transform (SIFT). It is used to detect the local features of an image in a very quick and reliable manner [10]. In SURF, first of all, the key-points of an image are perceived, and then related consistent descriptors are calculated [6].

3) Binary Robust Invariant Scalable Keypoints (BRISK): BRISK is a binary descriptor in which key-points are selected, and then a sampling pattern is applied to the neighbors of those key-points in an image. Every pair of pixels around the key-points is separated by two subsets, such as long-distance pair and short-distance pair [14].

D. Classification

Classification is important for the detection of fungal blast disease in rice crops. It imposes a class on the new sample with the help of learning from different classifier models by training [3]. Classification can be performed by using the actual image of the dataset or by using the features which have extracted. The main reason purpose of using classification is, it can detect plant disease automatically [9]. Classification with traditional classifiers can be done with the help of features. For the classification of rice crops, both convolutional and traditional classifiers have been used. All the feature values have been given as input to the below-mentioned classifiers by splitting 80% of data for training and 20% for testing.

1) Support Vector Machine (SVM): SVM is a supervised learning algorithm that uses Support Vector Classification (SVC) for classification purposes. It is a linear classification technique and has been found most competitive in machine learning algorithms for the classification of high-dimensional datasets [10]. SVM is easy to use and controls the complexity of decision and frequency error [20]. Equation (6) shows how the SVM classifier works at the backend. The accuracy achieved in the SVM classifier with the image size of 256 x 256 has the highest accuracy before PCA [6] as compared to other classifiers. The accuracy comparison of SVM with different sizes of the dataset is shown in Fig. 6.

$$argmax_{j=1\ldots M}\, g^j(x) \; where \; g^j(x) = \sum_{i=1}^{m} y_i a_i^j k(x, x_i) + b^j \quad (6)$$

2) Linear Discriminant Analysis (LDA): LDA is a supervised learning algorithm that finds the linear combination based on different features that can split two or more classes. It can also be used for dimensionality reduction purposes because it can be used for more than two classes for classification [21]. Like SVM, it is a linear classification technique. Equation (7) shows the discriminant for the linear variable, so this is the equation for the linear discriminant. The accuracy achieved in the LDA classifier before PCA for 256 x 256 image size is comparatively less than SVM classifier. The accuracy comparison for different image sizes is shown in Fig. 6.

$$\delta_k(k) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \quad (7)$$

TABLE I. FEATURES THAT HAVE BEEN USED

Sr. No.	*Features Type*	*Features Name*
1	Texture Features	Contrast, Correlation, Energy, Entropy, Homogeneity
2	Brisk Features	Scale, Orientation, Metric
3	Surf Features	Scale, Orientation, Metric

TABLE II. FORMULA FOR TEXTURE FEATURES

Eq. No.	*Features Type*	*Features Formula*
1	Contrast	$\sum_{i=1}^{n}\sum_{j=1}^{n}(i,j)^2 p(i,j)$
2	Correlation	$\sum_{i,j=1}^{n}\frac{p_{i,j}(i-\mu)(j-\mu)}{\sigma^2}$
3	Energy	$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(p(i,j)\right)^2$
4	Entropy	$\sum_{k=0}^{i=1} prk(log_2 prk)$
5	Homogeneity	$\sum_{i,j=1}^{n}\frac{p_{i,j}}{1+(i-j)^2}$

3) Logistic Regression (LR): Logistic regression is a statistical supervised machine learning algorithm that is used for classification purposes. It works based on the concept of probability, so it is also known as a predictive analysis algorithm. It uses the complex cost function known as 'Sigmoid function' instead of a linear cost function that is why sometimes it is not said as linear regression [23]. Equation (7) shows the complex cost function of logistic regression, and equation (8) is used for the multiple regression problems, which take more than one predictor. The results for multiple logistic are comparatively better than linear regression. The accuracy achieved in the LR classifier before PCA for an image size of 256 x 256 is smaller than both SVM and LDA classifiers. The accuracy comparison histogram for logistic regression classifier for different image sizes is given in Fig. 6.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases} \quad (8)$$

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (9)$$

4) Naïve Bayes (NB): Naïve Bayes is a probabilistic algorithm that works based on the Bayes' theorem. This classifier takes every feature conditionally independent with others [10]. With this assumption, it calculates the likelihood of the data using Bayes' theorem with the product of conditional probability [24]. The best hypothesis in Naïve Bayes theorem can be chosen based on equation (10). The accuracy achieved in the NB classifier is the lowest accuracy than all other classifiers. The accuracy comparison plots are given in Fig. 6.

$$\hat{y} = argmax\ P(y)\ \prod_{i=1}^{n}\left(P(x_i|y)\right) \quad (10)$$

5) Decision Tree (DT): The decision tree is the most useful classifier in machine learning algorithms because it takes the most suitable attribute at its root node [23]. It works based on the entropy and information gain approach for the construction of its tree. Equation (11) shows the formula for entropy, and equation (12) is for gain. If the entropy is more positive, then the instances will be more heterogeneous [24]. The accuracy achieved in the DT classifier for 256 x 256 image size is more than SVM and all other classifiers before PCA. The accuracy comparison histograms are given in Fig. 6.

$$E = \sum_{i=1}^{c} -p_i log_2 p_i \quad (11)$$

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \quad (12)$$

6) Random Forest (FR): Random forest is a supervised machine learning algorithm, mainly used for classification purposes. It has comparatively better accuracy results than that decision tree classifier because it works based on a decision tree [7] [9] [10]. Random forest is mostly used to avoid overfitting in decision tree classifiers. It constructs the trees which have been trained using the data samples training and features [10]. The accuracy achieved in the RF classifier for 256 x 256 image size before PCA is greater than all other classifiers. The accuracy comparison histograms are shown in Fig. 6. A random forest classifier has been concluded best for the fungal blast disease detection based on already defined features. The feature importance graph for random forest classifier is shown in Fig. 9, which shows that the Metric of BRISK features has the most importance in the random forest algorithm.

a) Performance of classification: The performance of all the above classifiers can be measured based on their classification report in terms of training and testing results [3]. The performance can be measured based on four parameters accuracy, precision, recall, and f1-score on the testing results. All these parameters are measured with true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) from the confusion matrix [3]. The formula for every parameter is shown in Table III. This table shows the formula for terms used in the classification report.The histogram is plotted only for f1-score because it is the combination of both precision and recall. The accuracy comparison of the f1-score for all classifiers is shown in Fig. 5.

TABLE III. CLASSIFICATION REPORT PARAMETERS FORMULA

Eq. No.	*Features Type*	*Features Formula*
13	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
14	Precision	$\frac{TP}{TP + FP}$
15	Recall	$\frac{TP}{TP + FN}$
16	F1-Score	$2 * \frac{Precision * Recall}{Precision + Recall}$

E. K-Fold Cross Validation

Cross-validation is a machine learning algorithms technique which mostly used to test the machine learning models are performing effectively. In the case of the limited dataset, the cross-validation can also be used as resampling to evaluate a model [14]. In this case, K-fold cross-validation has been performed on the training dataset taking 10 folds for the confirmation that all the created classifiers have not been overfitted [3]. In K-Fold the process repeats itself for k times so there can be k times Mean Square Error (MSE), and equation (13) shows the formula for MSE. All the accuracy results have been achieved with 10 fold cross-validation; the comparison histogram is shown in Fig. 6.

$$cv_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \quad (17)$$

F. Pricncipal Component Analysis (PCA)

PCA is an unsupervised machine learning algorithm used for dimension reduction in the case of a large number of dimensions or features. It shows better accuracy results after reducing the dimension of features of the original dataset because the models with high dimensions or a huge number of features can perform very slowly and most of the time fail to perform classification [6]. PCA is also used to remove the

overfitting in classifier models and it also improves the performance of model accuracy at a very low cost [20]. In this case, PCA is also applied because there are total 11 number of features and it is difficult for a model to make the decision so from these 11 only 6 PCA components have been taken for the classification purposes and the accuracy results for the 6 components are comparatively similar to the results before PCA. This proves that reducing the dimension or the number of features gives almost the same accuracy as without PCA. For comparison purposes, the computed accuracy for PCA 6 and 7 components, and all the comparison plots are given in the results section, shown in Fig. 7 and Fig. 8.

G. Transfer Learning

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task [27]. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in a skill that they provide on related problems. Transfer methods tend to be highly dependent on the machine learning algorithms being used to learn the tasks and can often simply be considered extensions of those algorithms [17]. In transfer learning, the initial steps of image acquisition and image preprocessing are the same as shown in Fig. 3, which are applied for traditional classifiers. Data augmentation is a strategy that enables us to significantly increase the diversity of data available for training models without actually collecting new data [23] [24]. In this process, the dataset has been divided into an augmented and unaugmented form which has been further passed for transfer learning techniques, such as cropping, padding, and horizontal flipping are used to augment the data to train a large neural network with small dataset.

H. VGG-16

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman [25], the model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. The model of VGG-16 is shown in Fig. 4, it includes 13 Convolutional layers, 5 pooling layers, and 3 dense/fully connected layers.

a) Convolutional layer: The Convolutional layer is the building block of the neural network; it is application of a filter to an input that results in an activation. A feature map is generated with the repeated application of the same filter in a map of activations, indicating the locations and strength of a detected feature in an input, such as an image [26].

b) Pooling layer: The pooling layer is placed right after the convolutional layer, it provides downsampling of feature maps by summarizing the presence of features in patches of the feature map. Average pooling and max pooling are two common methods that summarize the average presence of a feature and the most activated presence of a feature respectively [22][27].

c) Fully connected layer: Fully connected layers are an essential component of Convolutional Neural Networks (CNNs), which have been proven very successful in recognizing and classifying images for computer vision. The CNN process begins with convolution and pooling, breaking down the image into features, and analyzing them independently. The result of this process feeds into a fully connected neural network structure that drives the final classification decision [27].

d) Softmax/sigmoid layer: The Softmax function is sometimes called multi-class logistic regression because the softmax is a generalization of logistic regression that can be used for multi-class classification, whereas the sigmoid function is used for logistic regression or binary classification [27].

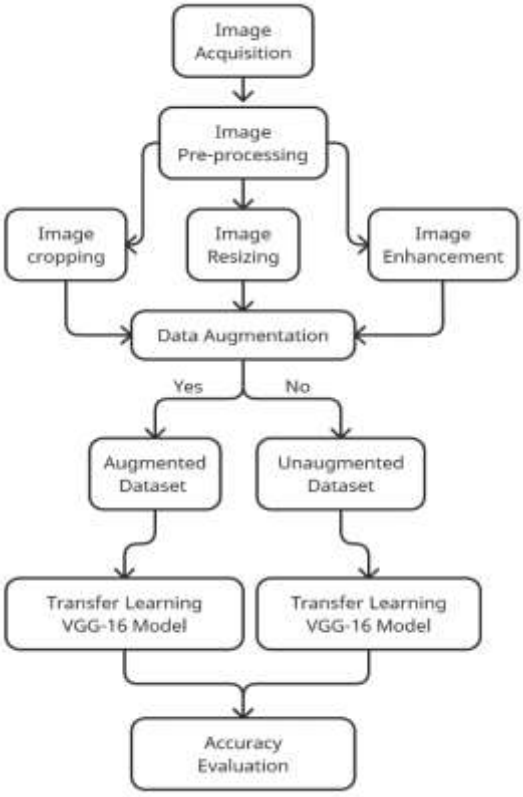


Fig. 3. Proposed Work Flow Chart for Transfer Learning.

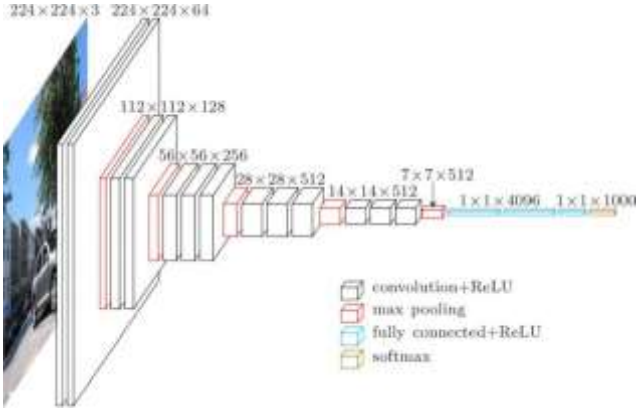


Fig. 4. VGG-16 Model which Includes 13 Convolutional, 5 Pooling and 3 Dense/Fully Connected Layers.

7) Regularization: Dropout is a regularization method that approximates training a large number of neural networks with different architectures in parallel [22]. During training, some number of layer outputs are randomly ignored or "dropped out." This has the effect of making the layer look-like and be treated-like a layer with a different number of nodes and connectivity to the prior layer. In effect, each update to a layer during training is performed with a different "view" of the configured layer [27]. The dropout layer is used in the transfer learning VGG16 model after the relu activation layers to randomly drop the weights and generalize better to remove the overfitting.

IV. Results and Discussion

In this section, all the results of classification accuracy and report have been discussed thoroughly. A comparative analysis has been taken for every classifier's performance and accuracy results before PCA and after PCA. All the resulting histograms are discussed in this section.

A. F1-Score Classification Analysis

The accuracy of all the above discussed classifiers is shown in Fig. 5. The given F1-score comparison plot is the average of both healthy and unhealthy rice crops. F1-score is the combination of precision and recall, so here, F1-score values for the SVM and random forest classifiers with an image size of 256 x 256 and 512 x 512 are higher than other classifiers. It proves that for the fungal blast disease in rice seed, the classification with SVM and the random forest is much better than other classifiers. The random forest has better results for the image size of 256 x 256. F1-score values for the Naïve Bayes classifier are lowest than other classifiers, but for the case of 256 x 256 image size, it has a high score. The remaining classifier has an almost related F1-score, so from this comparison histogram, it has been concluded that based on precision and recall results, the SVM and random forest, both classifiers have higher results and can perform better for the image size of 256 x 256.

B. Classification Accuracy Analysis before PCA

The accuracy comparison histogram before PCA is shown in Fig. 6. Based on the healthy and unhealthy dataset for the fungal blast disease, the final results are shown in the histogram conclude that the accuracy of SVM and random forest classifiers are higher than other classifiers. Six different classification models are applied for 10 folds cross-validation on 80% of the training and 20% of testing of the dataset. The classification has been performed on all image sizes of the dataset and it has been observed that the results for 256 x 256 image size are most accurate. The accuracy results that are achieved with 256 x 256 image size for 10-fold cross validation before PCA are given as under:

The SVM classifier has better performance with testing accuracy of 73.50%. Random Forest classifier also has best performance of 74.80% of testing accuracy. LDA classifier performed well with an accuracy of 70.55%, which is less as compared to SVM and random forest. Logistic regression classifier has an accuracy of 68.05%, which is less than LDA. The Decision Tree classifier has got an accuracy of 65.55%, which is smaller than logistic regression. While, Naïve Bayes achieved an accuracy of 62.33%, which is the lowest as compared to all classifiers, shown in Fig. 6. From the above classification results, Naïve Bayes classifier has very low accuracy while LDA and logistic regression have almost the same accuracies. The decision tree classifier also has good performance, but the results are smaller than SVM and random forest. Finally, it has been concluded that the SVM and random forest both classifiers have better performance with 10 folds cross-validation, and the accuracy is almost 73.50% and 74.80%.

C. Classification Accuracy Analysis after PCA

The PCA classifier is used to reduce the dimension or the features to get better results. In this case, PCA is also applied to reduce the number of features. PCA is applied with 10-fold cross-validation for 6 and 7 components with the reduction of 5 and 4 dimensions from 11 dimensions (features), shown in Fig. 4 and 5. From the comparison histogram before PCA, it can be seen that the results are good, but possibly due to the huge number of features, models get confused and did not perform well. So, here 6 and 7 PCA components are taken, which help the models to for a better decision. The comparison histogram after applying PCA is shown in Fig. 4 and 5. After applying PCA to every classifier similarly, 10 folds cross-validation has been applied for the removal of overfitting.

The accuracy results that are achieved with 256 x 256 image size for 10-fold cross validation for 6 PCA components are given as under: The SVM classifier has better performance with a testing accuracy of 69.03%. Random Forest classifier also has the best performance of 72.52% of testing accuracy. LDA classifier performed well with an accuracy of 68.55%, which is less as compared to SVM and random forest. Logistic regression classifier has an accuracy of 68.05%, which is less than LDA. The Decision Tree classifier has got an accuracy of 67.88%, which is smaller than logistic regression. While, Naïve Bayes achieved an accuracy of 65.53%, which is the lowest as compared to all classifiers, shown in Fig. 6. From the above classification results for 6 PCA components, it has been concluded that the testing accuracy for random forest classifier is higher than all others. So, for 6 PCA components, random forest classifier has better performance.

The accuracy results that are achieved with 256 x 256 image size, for 10-fold cross validation for 7 PCA components are given as under: The SVM classifier has better performance with testing accuracy of 71.45%. Random Forest classifier also has the best performance of 70.65% of testing accuracy. LDA classifier performed well with an accuracy of 68.67%, which is less as compared to SVM and random forest. Logistic regression classifier has a accuracy of 69.08%, which is less than LDA. The Decision Tree classifier has got an accuracy of 67.18%, which is smaller than logistic regression. While, Naïve Bayes achieved an accuracy of 66.12%, which is the lowest as compared to all classifiers, shown in Fig. 6. From the above classification results for 7 PCA components, it has been concluded that the testing accuracy for SVM classifier is higher than all others. So, for 7 PCA components, SVM classifier has better performance.

From overall results, it has been concluded that the accuracy for 6 and 7 components is almost similar to the accuracy for all components. It proves that after reducing the number of features in PCA almost the same results are achieved as before applying PCA. The dimensionality is reduced up to 4 and 5 features. So, both results before PCA and after PCA are almost the same and with the help of these traditional classifiers, maximum achieved accuracy is 75%.

D. VGG-16 Performance Analysis

The VGG-16 classifier is used to classify healthy and unhealthy images, it has been used in two conditions, without data augmentation and regularization, and with data augmentation and regularization. Both VGG-16 classifiers have top layers disabled, and a new model has been created using the pre-trained weights of VGG-16.

The results of VGG-16 without data augmentation and regularization are shown in Fig. 10, the validation accuracy is a maximum of 64%, and it has stopped learning, which indicates that the model is overfitting. Two techniques are used to reduce overfitting in the model, i.e., augmentation and dropout. In data augmentation, the data is increased artificially for the model to learn better in training epochs and regularization, and then dropout regularization is used in which the model randomly drops learned weights after every epoch, which helps the model not to become general. In Fig. 11, it can be seen that the validation accuracy of the model has increased to 71.28% after applying the data augmentation and dropout regularization technique. These techniques play a crucial part in the fine-tuning of the model to achieve the best results.

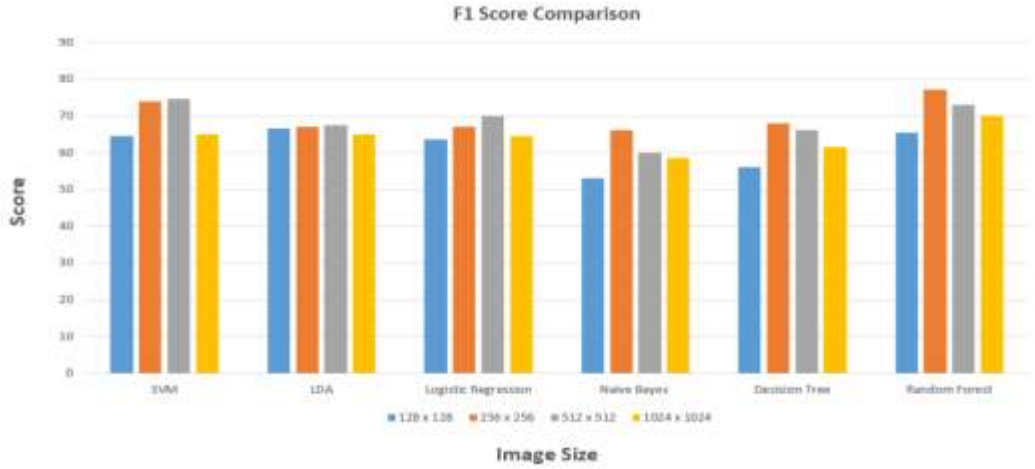


Fig. 5. Comparison Histogram of Every Classifier Discussed above for Classification Report Parameters, F1-Score which is Combinaion of Precision and Recall, this Histogram shws that F1-Score Values for SVM and Random Forest Classifier are Higher.

Fig. 6. Comparison Histogram of Every Classifier for Every Size of Image before PCA which Shows that SVM and Random Forest Classifiers have Higher Accuracy for the Image Size of 256 x 256.

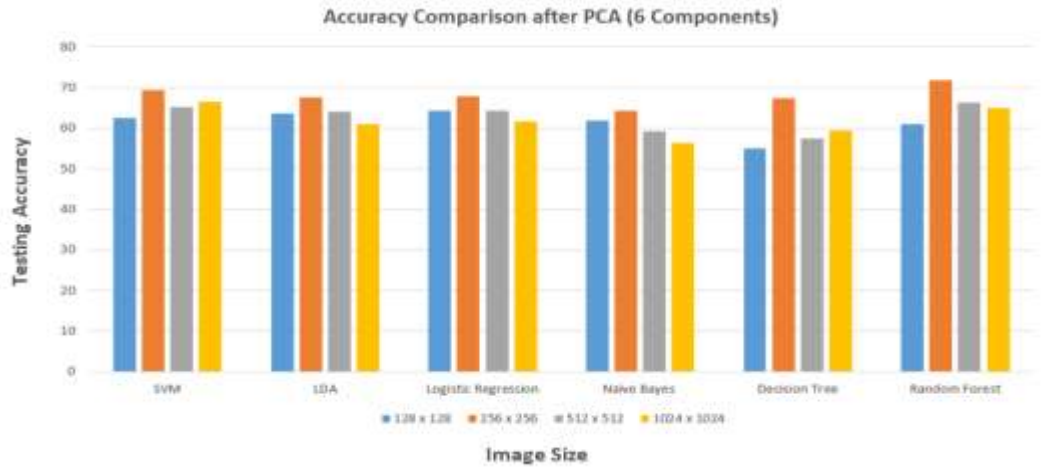


Fig. 7. Comparison Histogram of Every Classifier after PCA for 6 PCA Components which shows that SVM And random Forest Classifiers have Higher Accuracy for the Image Size of 256 x 256.

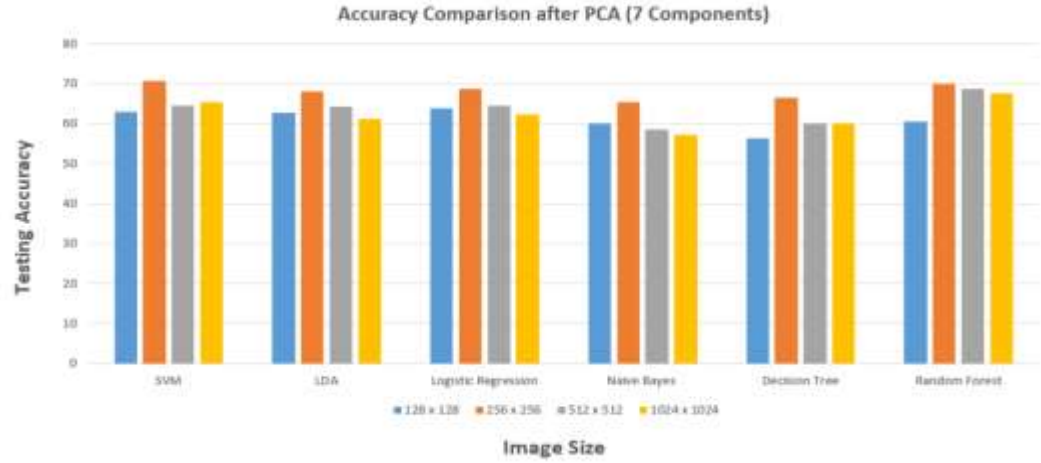


Fig. 8. Comparison Histogram of Every Classifier after PCA for 7 PCA Components which Shows that the Accuracy Results for SVM and Random Forest are Higher for Image Size of 256x256.

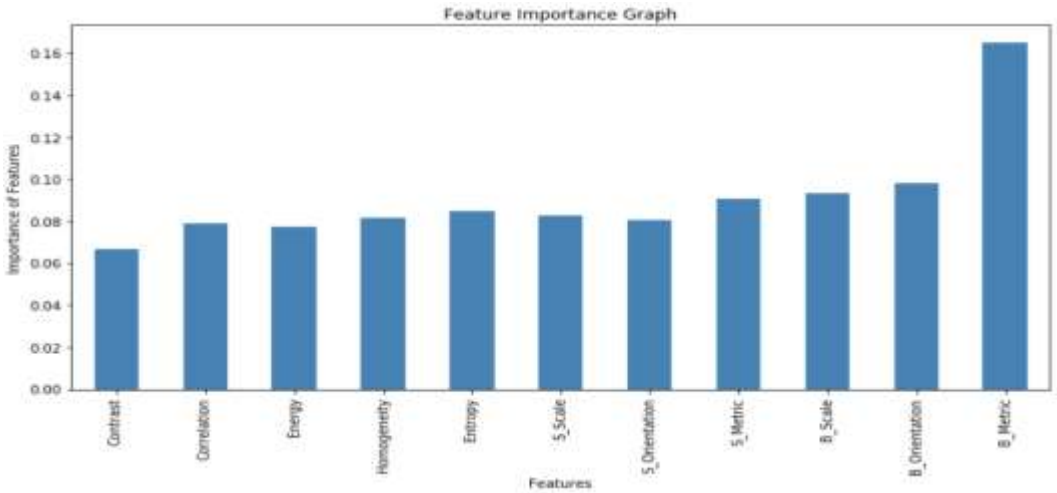


Fig. 9. Feature Importance Histogram for Random Forest Classifier where B and S in Features Axis Stands for BRISK and SURF which Shows that Metric of BRISK Features is the Most Important Feature in Random Forest Classifier.

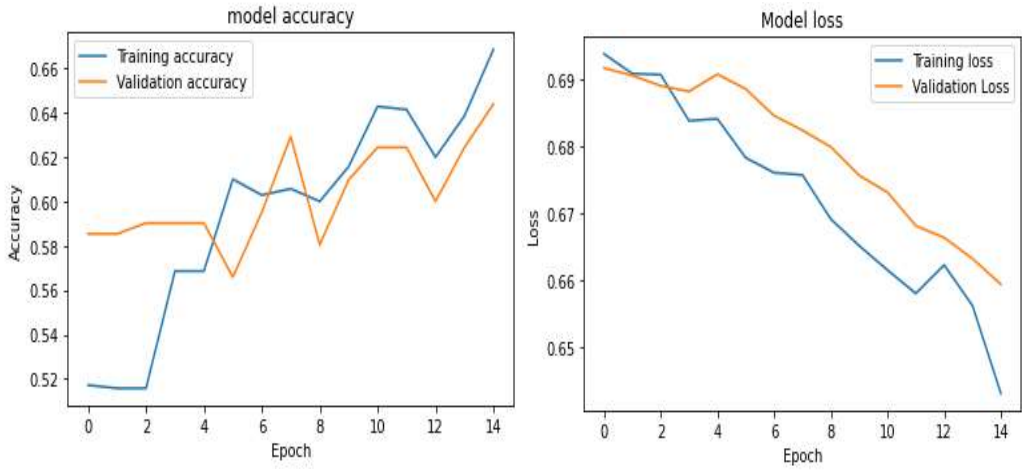


Fig. 10. Training Accuracy and Validation Accuracy along with Training Loss and Validation Loss of VGG-16 Model using Unaugmented Dataset and no Regularization.

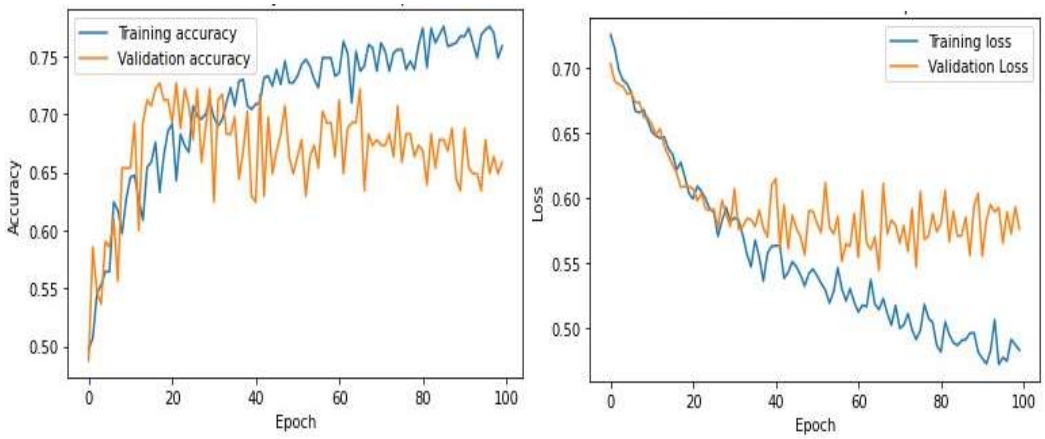


Fig. 11. Training Accuracy and Validation Accuracy along with Training loss and Validation Loss of VGG-16 Model using Augmented Dataset and Regularization.

V. Major Contribution

In this comprehensive research, the major contribution is the unique dataset of rice crops which has been collected from different fields of Kashmore, Pakistan. There are many publications for plant disease detection in general, but regarding the rice plant diseases, limited research work is done that is only for rice leaf diseases. In this research, the fungal blast disease has been detected on rice crop seed with different image processing techniques and machine learning algorithms, which is another major contribution.

VI. Conclusion

Plant disease detection plays an essential role in the growth of the economy and healthy crop production. In the proposed work, the fungal blast disease is detected in the seed of rice crop. This paper discussed the different image processing and machine learning techniques to detect fungal blast disease in rice crops. Image processing is used for the extraction of multiple features and extracted 11 different features from the models such as texture, SURF, and BRISK. As per this research, the mentioned features are beneficial for the detection of fungal blast disease, in which rice has brownish spots on its seed, shown in Fig. 2. In the machine learning portion, a comparative analysis regarding different machine learning algorithms based on disease detection with varying accuracies has been made. Seven different classifiers are used, including traditional and convolutional classifiers. After analyzing these traditional features and classifiers, the dataset has been used as input to transfer learning VGG-16 model, then trained the model with the unaugmented dataset and augmented dataset. After training, the validation accuracy of the trained model with the unaugmented dataset was 64%, while the accuracy of the trained VGG-16 model with the augmented dataset was 71.28%.

Finally, it has been concluded that after applying PCA with 10-fold cross validation, the random forest algorithm has still the best performance for the fungal blast disease detection with an accuracy of 73.12% for the testing dataset in the traditional classifiers whilst the highest accuracy from transfer learning dataset is of 71.28%. If analyzed, it is not a big difference as compared to the efforts that have been put in order to run the traditional classifiers while the images data was input to the transfer learning model.

VII. Future Scope

In future work, the plan is to develop a mobile application and an agricultural cultivating drone for fungal blast disease detection in rice crop seed during the field. This mobile application will help farmers to detect the disease in rice seed by capturing an image of the plant in the field, and they will get the most accurate and fast results on the spot. Similarly, an agricultural drone will visit the whole field and will monitor the plant's health. Based on those results of drone and mobile applications, the farmers can use related pesticides and fertilizers to improve the health of the crop. This technology will reduce the cost for extra use of pesticides, and farmers will get a good profit while giving only the needed pesticides to crops, it will be more beneficial for the economy of this country.

References

[1] S. D. Khirade and A. B. Patil, "Plant Disease Detection Using Image Processing," 2015 International Conference on Computing Communication Control and Automation, Pune, 2015, pp. 768-771, doi: 10.1109/ICCUBEA.2015.153.

[2] P. Panchal, V. C. Raman and S. Mantri, "Plant Diseases Detection and Classification using Machine Learning Models," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2019, pp. 1-6, doi: 10.1109/CSITSS47250.2019.9031029.

[3] U. B. Korkut, Ö. B. Göktürk and O. Yildiz, "Detection of plant diseases by machine learning," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404692.

[4] Raut Sandesh, Fulsunge Amit, Plant Disease Detection in Image Processing Using MATLAB, 2017, Volume-6, ISSN 2319-8753.

[5] O. Kulkarni, "Crop Disease Detection Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697390.

[6] L. S. Puspha Annabel, T. Annapoorani and P. Deepalakshmi, "Machine Learning for Plant Leaf Disease Detection and Classification – A Review," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0538-0542, doi: 10.1109/ICCSP.2019.8698004.

[7] K. Ahmed, T. R. Shahidi, S. M. Irfanul Alam and S. Momen, "Rice Leaf Disease Detection Using Machine Learning Techniques," 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/STI47673.2019.9068096.

[8] "Rice leaf diseases data set." https://archive.ics.uci.edu/ml/datasets/Rice+Leaf+Diseases. Accessed: 2019-09-27.

[9] Kurale, Neha & Vaidya, Madhav. (2018). Classification of Leaf Disease Using Texture Feature and Neural Network Classifier. 1-6. 10.1109/ICIRCA.2018.8597434.

[10] Anjna, Sood, M., & Singh, P. K. (2020). Hybrid System for Detection and Classification of Plant Disease Using Qualitative Texture Features Analysis. Procedia Computer Science, 167(2019), 1056–1065. https://doi.org/10.1016/j.procs.2020.03.404

[11] S. Ramesh *et al*., "Plant Disease Detection Using Machine Learning," *2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, Bangalore, 2018, pp. 41-45, doi: 10.1109/ICDI3C.2018.00017.

[12] P. Sharma, P. Hans and S. C. Gupta, "Classification Of Plant Leaf Diseases Using Machine Learning And Image Preprocessing Techniques," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 480-484, doi: 10.1109/Confluence47617.2020.9057889.

[13] R. Naga Swetha, V. Shravani, Monitoring of Rice Plant for Disease Detection using Machine Learning, Volume-9, ISSN 2249-8958, https://www.ijeat.org//papers/v9i3/C5308029320.pdf.

[14] Muhammad Kashif, Thomas M. Deserno, Daniel Haak, Stephan Jonas, Feature description with SIFT, SURF, BRIEF, BRISK, or FREAK? A general question answered for bone age assessment, Computers in Biology and Medicine, Volume 68, 2016, Pages 67-75, ISSN0010-4825, https://doi.org/10.1016/j.compbiomed.2015.11.006.

[15] M. A. Jasim and J. M. AL-Tuwaijari, "Plant Leaf Diseases Detection and Classification Using Image Processing and Deep Learning Techniques," 2020 International Conference on Computer Science and Software Engineering (CSASE), Duhok, Iraq, 2020, pp. 259-265, doi: 10.1109/CSASE48920.2020.9142097.

[16] Prajapati, Harshadkumar & Shah, Jitesh & Dabhi, Vipul. (2017). Detection and classification of rice plant diseases. Intelligent Decision Technologies. 11. 357–373. 10.3233/IDT-170301.

[17] U. Shruthi, V. Nagaveni and B. K. Raghavendra, "A Review on Machine Learning Classification Techniques for Plant Disease Detection," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 281-284, doi: 10.1109/ICACCS.2019.8728415.

[18] E. Hossain, M. F. Hossain and M. A. Rahaman, "A Color and Texture Based Approach for the Detection and Classification of Plant Leaf Disease Using KNN Classifier," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ECACE.2019.8679247.

[19] B. S. Kusumo, A. Heryana, O. Mahendra and H. F. Pardede, "Machine Learning-based for Automatic Detection of Corn-Plant Diseases Using Image Processing," 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Tangerang, Indonesia, 2018, pp. 93-97, doi: 10.1109/IC3INA.2018.8629507.

[20] L. S. Puspha Annabel, T. Annapoorani and P. Deepalakshmi, "Machine Learning for Plant Leaf Disease Detection and Classification – A Review," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0538-0542, doi: 10.1109/ICCSP.2019.8698004.

[21] Kumar, D. S. and Samrity. "Plant Species Identification using SIFT and SURF Technique." (2017), ISSN 2319-7064, https://www.ijsr.net/archive/v6i3/ART20171974.pdf.

[22] Arsa, D. M. S., & Susila, A. A. N. H. (2019). VGG16 in Batik Classification based on Random Forest. *Proceedings of 2019 International Conference on Information Management and Technology, ICIMTech 2019*, *1*(August), 295–299. https://doi.org/10.1109/ICIMTech.2019.8843844.

[23] Weiss, K., Khoshgoftaar, T. M., & Wang, D. D. (2016). A survey of transfer learning. In *Journal of Big Data* (Vol. 3, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-016-0043-6

[24] Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *ArXiv*.

[25] Khan, Ufaq. "Plant Disease Detection Techniques: A Review." (2019), IJCSMC (Vol. 8, No. 4), Pages 59-68, http://paper.researchbib.com/view/paper/206987.

[26] N.K Ambika, P Supriya, Detection of Vanilla Species by Employing Image Processing Approach, Procedia Computer Science, Volume 143, 2018, Pages 474-480, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.10.420.

[27] Sharma, P., Hans, P., & Gupta, S. C. (2020). Classification of plant leaf diseases using machine learning and image preprocessing techniques. Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering, 480–484. https://doi.org/10.1109/Confluence47617.2020.9057889

Investigation of Factors Affecting Employee Satisfaction of IT Sector

Eiman Tamah Al-Shammari
Kuwait University
College of Life Sciences
Shadadiya
Kuwait

***Abstract*—Job satisfaction or employee satisfaction has various definitions, but we can generalize it by how gratified an individual is with his or her job. Happy employees help to strengthen the company by lowering turnover and increasing loyalty. Job satisfaction also promotes a healthy working environment that helps to attract talent and increase productivity. However, little research has been done that focuses specifically on the IT sector. The goal of this research is to measure the level of satisfaction among Kuwaiti IT workers and discover tangible and intangible factors affecting their job satisfaction. To highlight factors contributing to positive satisfaction in the IT jobs in Kuwait, we propose a six-factor structural model, including compensation, workplace, intangible benefits, support, communication, and satisfaction. A targeted snowball descriptive survey was distributed via WhatsApp messages to Information Technology workers; 209 responses were collected after data cleaning. SPSS statistical software was used to analyze the data, with results indicating IT employees felt an average level of satisfaction. Additionally, several work-related variables were significantly associated with job satisfaction. Work position showed a statistically significant association with work satisfaction. Finally, individuals in a leading position reported higher satisfaction compared to individuals in non-leading positions.**

Keywords*—Job satisfaction; IT sector; productivity; intangible benefits; communication

I. Introduction

Job satisfaction determines how happy a person is with their job. Job satisfaction can have an immensely positive and negative effect on the workplace. Dissatisfied employees can decrease productivity and cause high turnover [1]. It can also enhance performance and affect customers' satisfaction directly and indirectly. In addition, job satisfaction is important as it can affect the quality of service provided to customers and affect customer retention [2].

A range of variables can affect the degree of job satisfaction of individuals. Pay and benefits, the perceived fairness of the promotion system, social relationships, upper management, job challenges, and job clarity are factors.

Previous studies have highlighted factors that lead to positive satisfaction, where other studies focused on exploring reasons behind dissatisfaction and turnovers. Factors were divided, into tangible including skills mismatch, commitment, gender differences, and stress [3,4,5]. Researchers classified these factors mainly into two categories: tangible and intangible factors. Tangible factors are simply those that can be quantified and measured such as salary, compensation, rewards, bonuses, work flexibility, training seminars, family or self-insurance, travel allowance, work environment, office location, office size, and promotion.

Whereas intangible factors are those of a qualitative nature. Hoppock defined intangible as the combination of psychological, physiological, and environmental circumstances that lead the worker to say I am satisfied [6]. Examples of such factors could be impressions, pressure, work relations, skills mismatch, commitment, flexible working hours, gender differences, stress, and feeling secure [3,4,5,7,8,9,10,11]. Promoting Ethical work standards is also considered an intangible factor [12]. Additionally, fairness of treatment can also be considered one [13].

It is hard to measure, yet we all differ in nature, and just as tangible benefits could be crucial to some employees, intangible factors could be more important to others, especially in an economy where there are a lot of cutbacks or layoffs.

Prior to moving forward with our study, we conducted a review of the related research conducted over the past twenty years. The next section will summarize these studies. Based on the findings we designed our survey with consideration of cultural differences.

The collection of the literature was directed towards the IT sector, as we lack such studies in Kuwait. As information technology departments became the backbone of every company, it became hard to find any organization that does not have an IT department. If we are allowed to generalize, as using technology became a required skill for every worker, we might consider all workers as IT workers.

This study would like to contribute and enrich studies in that subject, in a middle eastern country such as Kuwait. The concluding points will help decision-makers in improving workplace environments.

The following sections are arranged as follows: Section 2 will visit various previous studies focusing on job satisfaction for IT workers. Our approach will be discussed in Section 3. In Section 4, our results will be given. And finally, in Section 5, we will sum up our findings.

II. Literature Review

A. Defining Job Satisfaction

Hoppock defined job satisfaction as "any combination of psychological, physiological and environmental circumstances that cause a person to truthfully say I am satisfied with my job" [11]. Yet, the most widely used definition of job satisfaction was made by Locke, who defines it as "a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences" [14]. According to Vroom, job satisfaction is positive feedback from individual workers towards their current job [15]. Wanous and Lawler state job satisfaction as the "sum of job facet satisfaction across all facets of a job" [16]. This is very similar to Spector, who defines it as "how people feel towards their job from different aspects" [17] and Schermerhorn," as the emotional response towards various aspects of an employee's work" [18]. More definitions supported the same meaning.

Reilly describes job satisfaction as the feeling that a worker has towards his job, influenced by the perception of one's job [19]. Mansoor, Muhammad, Fida, Nasir, and Ahmad suggested a similar definition: how positively people feel about their job [20]. Ellickson and Logsdon defined job satisfaction as the degree to which employees like their work [21].

Phillips and Connell defined it as "the degrees to which employees are content with the job they perform" [22]. More attempts to define the concept of satisfaction have resulted in the definition being the final state of the psychological process [23]. Many studies have suggested many definitions, with the majority focusing on how the employee feels about his job in general.

B. Job Satisfaction Factors

There are, according to Arnold and Feldman, a number of factors that make people feel positive or negative about their jobs [24]. Researchers have contributed heavily to prioritize these factors based on their influence on job satisfaction.

Nwagwu conducted a Nigerian study to observe job satisfaction among IT artisans. The study's main discovery showed 300 IT artisans surveyed were dissatisfied with their jobs; however, high expectations of a breakthrough and the trend of IT were key reasons for staying in their jobs [25].

However, other studies have shown that financial factors and promotions are the leading factors for job satisfaction [6,26,27,28]. Studies have shown that low financial income leads to high insecurity [29]. In addition to financial factors, Akbar et al. explored additional factors such as prospects for the working environment, training, career growth and improvement [28].

Frontczak and Else focused on the indoor work environment's quality and building design, defining a good workplace space as when workers are granted a private office space with windows close by [8]. Lottrup, Stigsdotter, Meilby, and Claudi supported this claim in their research, empathizing on the importance of having buildings with green surroundings and window views [10].Additional factors such as flexible working hours, work relations, family insurance, allowance, promotion, and benefits were discussed by Alam and Shahi [9]. In addition, they highlighted the significance of positive reviews from an employee's superior. Other researchers found that work relationships and higher morale significantly influence the level of satisfaction [30,31]. Furthermore, high ethical expectations in the workplace lead to greater satisfaction [12]. Additional studies have concentrated on gender and how it can play an important role in work satisfaction [32, 33]. However, other studies have denied this claim [20, 34]. Kowal and Roztocki have argued that women are less satisfied with their jobs [35].

A study by Clark discovered that although females occupy a lower position in their average job and get lower income than their male counterparts, the expectations of females have been contended to be lower in comparison to males. Therefore, females tend to report greater job satisfaction levels [36].

A study by Bordin, Carina, Bartram, and Casimir conducted in Singapore amongst IT workers shows that psychological empowerment can increase job satisfaction and organizational commitment. Additionally, the study revealed that similarly supervisory support is an important factor for the same reasons [37].

When examining other factors, other studies revealed that employees with flexible working hours had been seen to have higher job satisfaction than those without [9]. They tend to have more time in their private lives and harmonize with their profession [9]. They also found that forcing ethical work standards increased job satisfaction [14].

Lim discovered that wage, degree, a sense of belonging, faith in wanting to belong, a feeling of acceptance, job autonomy, and promotion opportunities were related to job satisfaction while evaluating it for library Information Technology staff [38].

Lumley, Coetzee, Tladinyane, and Ferreira carried out a cross-sectional analysis on a group of IT workers in companies in South Africa to investigate the connection between job satisfaction and employee organizational commitment. It suggested a significant relationship between job satisfaction and affective and normative commitment [39].

A study was conducted in India on IT workers has concluded that there is a strong link between job satisfaction and employee loyalty. And the main determinants of job satisfaction and employee loyalty are supervisory support, career growth, and job security [40].

Another research conducted in Singapore showed that personal accomplishment intercedes the relationship between emotional intelligence and job satisfaction for IT workers [41].

Wong, in a study from Hong Kong, argued that the effects of organizational culture on knowledge sharing leads to job satisfaction, which leads to an improvement in organization performance [42].

To Kumar, Roshan, Yashu, and Saran, Technostress leads to job dissatisfaction causing reduced productivity, high turnover, absenteeism, and poor performance, leading to job dissatisfaction and then lower organizational satisfaction [43].

Another study conducted by Adebiaye found that work attitude, cordial working relationships, and management support affect job satisfaction [44].

Sunil Misra and Kailash B. L. Srivastava, found that team building between bank employees generates competencies that positively affect employee effectiveness and job satisfaction [45].

Spann designed a study to investigate the relationship between the conflict and ambiguity role and job satisfaction for non-managerial IT. They concluded that there is a direct relationship of job satisfaction with both role conflict and role ambiguity [46].

Indian research conducted on IT professionals examined the relationship between work exhaustion and job satisfaction and discovered a negative correlation, additionally a positive correlation between work exhaustion and turnover intention. The study also considers the impact of emotional dissonance, role ambiguity, role conflict, the fairness of rewards, autonomy, and the perceived workload on IT professionals [47].

III. METHODS

A survey of a descriptive nature was used [48] to achieve our study goals, answering the following questions:

RQ1: What is the average job satisfaction score for Kuwaiti IT workers?

RQ2: What are the tangible and intangible factors influencing Kuwaiti IT workers job satisfaction?

RQ3: Which job characteristics are significantly associated with job satisfaction?

A targeted snowball survey was distributed via WhatsApp to Information technology workers, of which 209 responses were collected after data cleaning. The survey contained five-part sections completed by all respondents. The first-part is the demographic questions that consists of four questions, followed by the job characteristics the job characteristics which comprises of seven questions. The next sections were organized as follows: tangible benefits, intangible benefits, work relations questions, and general satisfaction related questions.

Independent variables were conceptualized within five domains: 1) Compensation, 2) workplace, 3) intangible benefits, 4) work relations, and 5) support. Job satisfaction is considered a dependent variable.

Continuous variables were summarized using means and standard deviations and categorical variables such as demographic and work characteristics were summarized using counts and percentages.

Histograms were used to assess the presence of univariate outliers. Scaled variables were also examined for points above or below three standard deviations from the mean. Data was explored for missing observations prior to the analysis. Histograms were also inspected for normality. Mahalanobis distance was used to check for multivariate normality.

Exploratory factor analysis was performed using maximum likelihood. Oblimin rotation (with Kaiser Normalization) was used. Variables were removed if they loaded on more than 1 latent variable (>0.4 on more than 1 latent variable) or did not load significantly on any of them (< 0.5).

Confirmatory factor analysis was performed to assess whether the data fit the hypothesized measurement model previously defined. Six, five, and four factor solutions were tested to assess the most appropriate factor structure to use. Reliability of the constructs was assessed using Cronbach's alpha and composite reliability. A value greater than 0.7 was considered satisfactory. The convergent validity of the constructs was assessed using the average variance extracted which should be greater than 0.5 for all constructs. Divergent validity was assessed by comparing the correlations between latent variables to square root the average variance extracted $\sqrt{AVE}$. Divergent validity was met if none of the correlations between latent variables was higher than square root the AVE. Individual indicators were allowed to load on only one factor and the latent variables were allowed to freely co-vary. The overall model fit was assessed using the following indices:

- C_{min}/df.
- The root mean square error of approximation (RMSEA) and the corresponding 90% Confidence interval.
- The Tucker–Lewis index (TLI).
- The comparative fit index (CFI).
- The standardized root mean square residual (SRMR).

The lower bound of good fit for the TLI and the CFI is considered to be 0.90. For the RMSEA and the SRMR, the upper bounds for good fit are considered to be 0.08 and 0.10, respectively. C_{min}/df less than 5 was considered an indication of good model fit (Table I). These cut off criteria for model fit were used as previously defined [49].

Hypotheses were tested using structural equation modelling (SEM).

Scale reliability analysis was performed using Cronbach's alpha. Cronbach's alpha is a measure of internal consistency which assesses how closely related a set of items are as a group. Cronbach's alpha is a function of the number of test items and the average inter-correlation among the items. The acceptable value for Cronbach's α is > 0.7.

TABLE I. THRESHOLD TO IDENTIFY GOOD MODEL FIT

Measure	**Threshold**
X2/df (C_{min}/df)	<3 good, < 5 acceptable
TLI	>0.95 excellent, > 0.9 good
CFI	>0.95 excellent, > 0.9 good
SRMR	< 0.08
RMSEA	< 0.05 good, 0.05 – 0.1 moderate
RMSEA 90% CI	< 0.1

SEM was performed to assess the association of the independent latent variables with the main DV (satisfaction with work). Model fit was assessed using the same previously mentioned fit measures. The R^2 was also calculated for the DV. R^2 represents the proportion of variance in the DV that is explained by IVs. Hypothesis testing was performed at 0.05 significance level.

Standardized coefficients were used to compare the effects of the independent variables included in the SEM. The standardized coefficients divide the size of the effect by the relevant standard deviations. So instead of being in terms of the original units of X and Y, the standardized regression coefficients are in terms of standard deviations which facilitates comparing regression coefficients. The R^2 is the squared multiple correlation and was used to assess the proportion of variance in the dependent variables that is explained by the independent variables. Statistical analysis was performed using SPSS v25 and R studio v1.1.463.

A. Satisfaction Across Kuwaiti IT Workers

Means and standard deviations were used to summarize the distribution of job satisfaction across various demographic and work characteristics. Scores for latent variables were computed by averaging the scores for the items included in the final CFA and SEM. One-way ANOVA was used to assess the association of various demographic and work factors with job satisfaction. One-way ANOVA was used since the DV (job satisfaction) is continuous in nature. Moreover, it can accommodate IVs with two or more levels unlike independent t-test which can only accommodate IVs with only two levels.

IV. Results

The initial data included 218 responses (n = 218). Nine responses were identified as outliers using Mahalanobis distance and were removed from the analysis (n = 209). Table II shows the characteristics of the study sample.

Table III shows the final factor structure. Six factors were identified: compensation (2 variables), workplace (6 variables), intangible benefits (2 variables), communication (2 variables), support (3 variables), and satisfaction (3 variables).

After excluding variables that did not meet the criteria, 18 items were used in the final analysis. These items formed a six-factor structure and none of the items loaded on more than a factor (latent variable).

A. Confirmatory Factor Analysis Results

1) Model choice: Results for CFA show that the six-factor solution provided an appropriate fit for the data. Workplace and compensation were used as two separate latent variables although both of them represent one aspect of the tangible benefits. This was done since model fit showed that combining them as one latent variable (five-factor model 1) resulted in poor model fit compared to the six-factor structure. Poor fit was also observed when communication and support were forced to load as one latent variable (five-factor model 2).

Results show that the six-factor model fits the data better compared to all remaining models as indicated by the AIC, and RMSEA. The TLI and CFI were also higher for the six-factor model. Likelihood ratio test showed that the six-factor model was significantly better compared to the remaining three models (Table IV). Thus, the six-factor solution was deemed appropriate since all fit measures were within the acceptable range. In addition, the C_{min}/df and the SRMR were 0.511 and 0.05 for the six-factor model, respectively.

TABLE II. Descriptive Statistics for the Study Sample

		Count	%
Age	20-25	16	7.7%
	26-31	65	31.1%
	32-37	85	40.7%
	38+	43	20.6%
Gender	Male	102	48.9%
	Female	107	51.1%
Education	High school or equivalent	29	13.9%
	Bachelor degree	125	59.8%
	Graduate	55	26.3%
Marital status	Single	33	15.8%
	Married	133	63.6%
	Divorced or separated	35	16.7%
	Widowed	8	3.8%
Income (month)	less than 700 KD	8	3.8%
	700 to less than 1000 KD	40	19.1%
	1000 to less than 1300 KD	75	35.9%
	1300 or more	86	41.1%
Work	Public Sector	113	54.1%
	Privet Sector	66	31.6%
	Mixed	30	14.4%
Position	A leading position	60	28.7%
	Non- leading position	149	71.3%
Experience at current job	Less than one year	6	2.9%
	1-5 years	65	31.1%
	5-10 years	78	37.3%
	More than 10 years	60	28.7%
Prior jobs	This is my first job	60	28.7%
	1	90	43.1%
	2+	59	28.2%
Relations at current job	Yes	67	32.1%
	No	142	67.9%
Job close to home	Yes	57	27.3%
	No	99	47.4%
	Somewhat	53	25.4%

TABLE III. PATTERN MATRIX FOR THE FINAL ROTATED FACTOR SOLUTION

	Factor					
	Cm	**SAT**	**WP**	**SP**	**NT**	**CP**
I am compensated for my hard work						0.537
I am satisfied with the benefits and payments made by my company						0.697
Comfortable office furniture positively affects my performance			0.704			
The color of the furniture affects my mood			0.701			
I feel more comfortable in a private office			0.804			
My office window view increases my productivity			0.800			
A clean workplace increases my performance			0.677			
Office space positively impacts my performance			0.618			
My current job matches my skills					-0.711	
My job takes advantage of my skills and abilities					-0.601	
I am encouraged when I have a good communication with my superiors	0.519					
Good communication between me and my colleagues increases my productivity	0.919					
I am receiving enough support from my supervisors / managers				-0.638		
My supervisor clearly identifies my daily responsibilities				-0.676		
My officials provide regular feedback on my performance				-0.982		
I am associated with my work		0.768				
I'm never considering leaving my current job		0.944				
In general, I am satisfied with my work		0.721				

Extraction Method: Maximum Likelihood.
Rotation Method: Oblimin with Kaiser Normalization.
CM: Communication, Sat: Satisfaction, WP: Workplace, SP: Support, NT: Intangible benefits, CP: Compensation

TABLE IV. CONFIRMATORY FACTOR ANALYSIS FOR VARIOUS MODELS

Model	Df	AIC	CFI	TLI	RMSEA	LR test X^2 (P)
Six-factor model	51	9466	0.952	0.939	0.076	-
Five-factor model 1	46	9554	0.922	0.904	0.096	97.85 (< 0.001)
Five-factor model 2	46	9764	0.853	0.82	0.131	308 (< 0.001)
Four-factor model	42	9845	0.825	0.792	0.141	397 (< 0.001)

Four factor model: Tangible, intangible benefits, communication, satisfaction.

B. Convergent and Divergent Validity

Results show that reliability was acceptable for all constructs (~0.7 or higher for all constructs). Convergent validity was confirmed by the fact that AVE was greater than 0.5 for all constructs (Table V). Divergent validity was assessed by comparing $\sqrt{AVE}$ of the construct to the correlation with the remaining latent variables ($\sqrt{AVE}$ should be higher than any corresponding correlation). This assumption was met for all constructs except for workplace that showed a strong correlation with communication (0.89). Factor loadings were greater than 0.7 for all variables (Fig. 1).

C. Structural Equation Modelling

A structural model was assessed in which satisfaction was used as the DV while all remaining five constructs were used as IVs. The proposed structural model (Fig. 2) was a good fit for the data as shown by CFI (0.964), TLI (0.954), RMSEA (0.065), upper 90% RSMEA confidence interval (0.078), and SRMR (0.054). All the proposed relations were statistically significant (Table IV).

Results show that the five IVs explain 72.1% of the variance in the DV (satisfaction of IT workers) as shown by the R^2. All five variables showed a statistically significant association with satisfaction with work. Compensation showed a statistically significant positive association with satisfaction (Std. β = 0.263, P < 0.05). This means that satisfaction increases by 1 standard deviation (SD) for each 1 SD increase in compensation which indicates that IT workers are more likely to be satisfied with work if they report satisfaction with payment. Effect of workplace showed a statistically significant negative association with job satisfaction (Std. β = -0.433, P = 0.002). This indicates that workers who are more affected by the workplace are less likely to be satisfied with the job.

Intangible benefits showed a statistically significant positive association with job satisfaction (Std. β = 0.413, P < 0.001). A similar result was observed with communication (Std. β = 0.323, P = 0.019) and support (Std. β = 0.278, P = 0.003). These results indicate that better communication with co-workers, support, as well as intangible benefits are associated with higher satisfaction with work. Comparing the standardized coefficients show that intangible benefits were the

strongest positive influencing factor. For each 1 SD increase in intangible benefits, satisfaction with work increases by 0.413 SD.

D. Job Satisfaction among Kuwaiti IT Employees

The average satisfaction with work was 3.04 (1.04) among Kuwaiti-IT workers which indicates a neutral state of satisfaction among the IT employees (Table VI). Table VII shows that several work-related variables were significantly associated with job satisfaction. Position showed a statistically significant association with satisfaction with work (F = 3.514, P < 0.1). Individuals in a leading position reported higher satisfaction compared to individuals in non-leading positions (3.26 vs. 2.96).

Number of previous jobs showed a statistically significant association with satisfaction (F = 5.47, P < 0.05). The mean satisfaction score was also lower among participants with two or more previous jobs compared to individuals who had 1 previous job or less (2.78 vs. 3.1). Job location also showed a statistically significant association with satisfaction (F = 4.987, P < 0.05). Individuals who reported having a job near home reported higher satisfaction compared to those who did not (3.3 vs. 3).

TABLE V. CORRELATION, DIVERGENT AND CONVERGENT VALIDITY FOR LATENT CONSTRUCTS

Model	α	AVE	CP	WP	NT	CM	SP	SAT
CP	0.73	0.58	0.76					
WP	0.93	0.69	0.63	0.83				
NT	0.69	0.77	0.69	0.63	0.88			
CM	0.77	0.84	0.62	0.89	0.678	0.92		
SP	0.84	0.69	0.67	0.43	0.738	0.518	0.83	
SAT	0.76	0.76	0.67	0.41	0.75	0.53	0.73	0.87

AVE: Average variance extracted

$\sqrt{AVE}$ is shown on the diagonal in bold

CM: Communication, SAT: Satisfaction, WP: Workplace, SP: Support, NT: Intangible benefits, CP: Compensation.

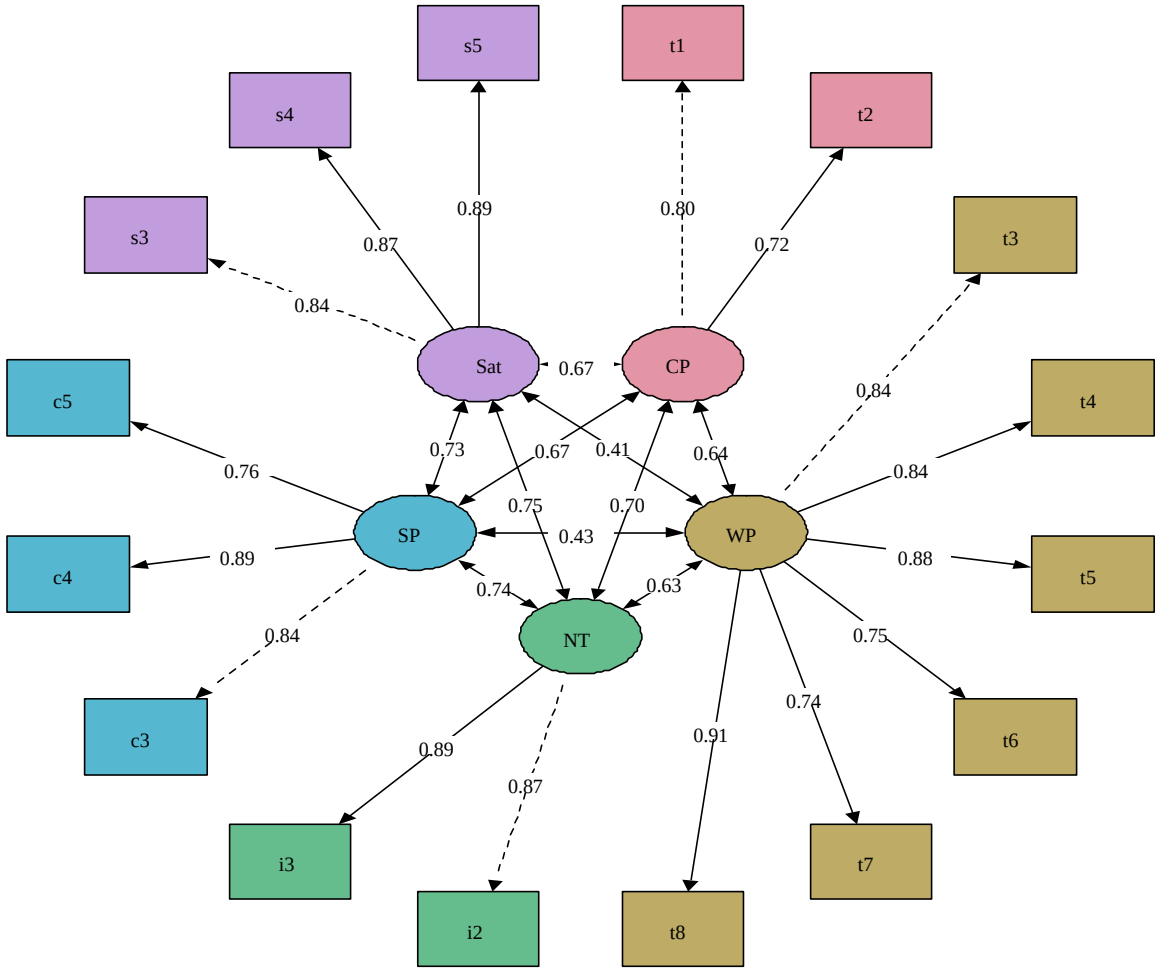


Fig. 1. Confirmatory Factor Analysis Results.

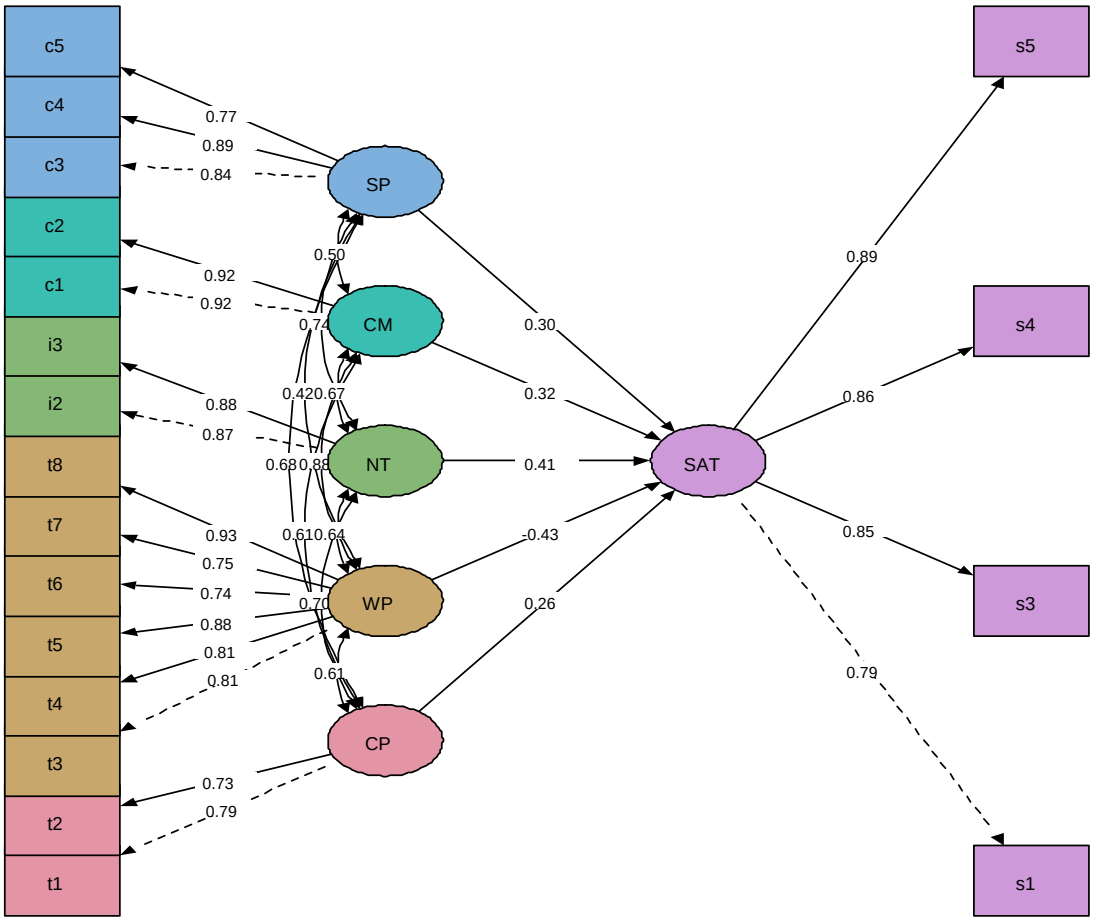


Fig. 2. Proposed Structural Model.

TABLE VI. STRUCTURAL MODEL ANALYSIS RESULTS ($R^2 = 0.721$)

IV	β	Std. β	SE	Z	P
CP	0.225	0.263	0.095	2.376	0.017*
WP	-0.354	-0.433	0.114	-3.118	0.002*
NT	0.352	0.413	0.096	3.661	< 0.001*
CM	0.241	0.323	0.102	2.355	0.019*
SP	0.278	0.3	0.092	3.008	0.003*

Satisfaction was used as the dependent variable in the model

CM: Communication, SAT: Satisfaction, WP: Workplace, SP: Support, NT: Intangible benefits, CP: Compensation

TABLE VII. DESCRIPTIVE STATISTICS FOR LATENT VARIABLES

Latent variable	Score
CP	2.72 (1.15)
WP	3.39 (1.15)
NT	3.14 (1.14)
CM	3.51 (1.24)
SP	2.91 (1.01)
SAT	3.04 (1.04)

CM: Communication, SAT: Satisfaction, WP: Workplace, SP: Support, NT: Intangible benefits, CP: Compensation

TABLE VIII. SATISFACTION ACROSS VARIOUS DEMOGRAPHIC AND WORK CHARACTERISTICS

		Mean	SD	F	P
Age	20-25	3.31	1.30	0.514	0.673
	26-31	2.96	0.97		
	32-37	3.07	0.96		
	38+	3.03	1.19		
Gender	Male	2.97	1.07	0.92	0.339
	Female	3.11	1.01		
Education	High school or equivalent	2.89	1.12	0.45	0.638
	Bachelor degree	3.05	1.05		
	Graduate	3.11	0.98		
Marital status	Single	2.95	1.10	0.821	0.483
	Married	3.00	1.06		
	Divorced or separated	3.26	0.87		
	Widowed	3.29	1.09		
Income (month)	less than 700 KD	3.08	0.99	0.757	0.52
	700 to less than 1000 KD	2.83	1.12		
	1000 to less than 1300 KD	3.05	0.95		
	1300 or more	3.13	1.08		
Work	Public Sector	3.05	1.16	0.077	0.926
	Privet Sector	3.07	0.82		
	Mixed	2.98	1.00		
Position	A leading position	3.26	1.08	3.514	0.062#
	Non- leading position	2.96	1.01		
How long have you been at this job	Less than one year	3.83	1.83	1.614	0.187
	1-5 years	3.12	0.97		
	5-10 years	3.02	0.99		
	More than 10 years	2.92	1.06		
Previous jobs	<2	3.15	1.04	5.47	0.02*
	2+	2.78	1		
Nearly relation at current job	Yes	3.17	1.07	1.537	0.216
	No	2.98	1.02		
Job near home	Yes	3.3	1.04	4.987	0.027*
	No/Somewhat	3	1		
# P < 0.1, * P < 0.05					

V. CONCLUSION

Job satisfaction is one of the main challenges facing the administration of all organizations. The average satisfaction score in the current analysis indicates a moderate level of satisfaction for Kuwaiti IT workers. The proposed six-factor structural model (compensation, workplace, intangible benefits, support, communication and satisfaction) was a good fit for the data as indicated by fit measures, convergent and divergent validity. Analysis results supported the pre-defined hypotheses. Compensation (tangible benefits), communication, support, intangible benefits showed a statistically significant positive association with job satisfaction. Higher levels of these variables result in higher job satisfaction. The perception of workplace (tangible benefits) showed a statistically significant negative association with job satisfaction. Individuals who are more affected by the workplace environment were less likely to report job satisfaction which supports the association between workplace and job satisfaction. The five IVs explained 72.1% of the variance in the DV (job satisfaction).

Our findings suggest that managers need to review current pay policies in order to build a satisfactory working atmosphere and offer fair pay, provide clear job instructions, and facilitate positive co-worker relationships.

Three characteristics related to work have shown a statistically significant association with job satisfaction: job position, number of previous jobs and location of work. Participants in a leading position are more likely to be satisfied with the job than those who are not. Participants with two or more previous jobs were less likely to be satisfied with the job than those with one or less previous job. Finally, participants who work in a job near their home were more likely to be satisfied than others who live far from their work.

As the present study was confined to participants working in IT field, it is not possible to generalize the findings to other professional contexts and regions. Furthermore, the sample is very narrow with limited factors, including more factors and a broader sample, to be considered in future studies.

REFERENCES

[1] Stone JR, Lewis MV. “College and career ready in the 21st century: Making high school matter”. Teachers College Press; 2012 Apr 6.

[2] Zeytinoglu, Işık U., Aşkın Keser, Gözde Yılmaz, Kıvanç Inelmen, Arzu Özsoy, and Duygu Uygur. "Security in a sea of insecurity: job security and intention to stay among service sector employees in Turkey." The International Journal of Human Resource Management 23, no. 13 (2012): 2809-2823.

[3] Rayton, Bruce A. "Examining the interconnection of job satisfaction and organizational commitment: An application of the bivariate probit model." The International Journal of Human Resource Management 17, no. 1 (2006): 139-154.

[4] Christen, Markus, Ganesh Iyer, and David Soberman. "Job satisfaction, job performance, and effort: A reexamination using agency theory." Journal of marketing 70, no. 1 (2006): 137-150.

[5] Cohrs, J. C., Abele, A. E., & Dette, D. A. (2006). Integrating situational and dispositional determinants of job satisfaction: Findings from three samples of professionals. The Journal of Psychology, 140(4), 363−395.

[6] Hoppock, Robert. "Job satisfaction." (1935).

[7] Parvin, Mosammod Mahamuda, and MM Nurul Kabir. "Factors affecting employee job satisfaction of pharmaceutical sector." Australian journal of business and management research 1, no. 9 (2011): 113.

[8] Frontczak, Monika, Stefano Schiavon, John Goins, Edward Arens, Hui Zhang, and Pawel Wargocki. "Quantitative relationships between occupant satisfaction and satisfaction aspects of indoor environmental quality and building design." Indoor air 22, no. 2 (2012): 119-131.

[9] Alam, Shahi Md Tanvir. "Factors affecting job satisfaction, motivation and turnover rate of medical promotion officer (MPO) in pharmaceutical industry: A study based in Khulna city."Asian Business Review 1, no. 2 (2012): 126-131.

[10] Lottrup, Lene, Ulrika K. Stigsdotter, Henrik Meilby, and Anne Grete Claudi. "The workplace window view: a determinant of office workers’ work ability and job satisfaction." Landscape Research 40, no. 1 (2015): 57-75.

[11] Danish, Rizwan Qaiser, and Ali Usman. "Impact of reward and recognition on job satisfaction and motivation: An empirical study from Pakistan." International journal of business and management 5, no. 2 (2010): 159.

[12] Kowal, Jolanta, and Narcyz Roztocki. "Do organizational ethics improve IT job satisfaction in the Visegrád Group countries? Insights from Poland." Journal of Global Information Technology Management 18, no. 2 (2015): 127-145.

[13] Adams, J. Stacy. "Inequity in social exchange." In Advances in experimental social psychology, vol. 2, pp. 267-299. Academic Press, 1965

[14] Locke, Edwin A., David Sirota, and Alan D. Wolfson. "An experimental case study of the successes and failures of job enrichment in a government agency." Journal of Applied Psychology 61, no. 6 (1976): 701.

[15] Vroom, Victor Harold. "Work and motivation." (1964)

[16] Wanous, John P., and Edward E. Lawler. "Measurement and meaning of job satisfaction." Journal of applied psychology 56, no. 2 (1972): 95.

[17] _Spector, Paul E. Job satisfaction: Application, assessment, causes, and consequences. Vol. 3. Sage, 1997.

[18] Schermerhorn Jr, John R. "Management for productivity." (1984).

[19] O'Reilly III, Charles A. "Organizational behavior: Where we've been, where we're going." Annual review of psychology 42, no. 1 (1991): 427-458.

[20] Mansoor, Muhammad, Sabtain Fida, Saima Nasir, and Zubair Ahmad. "The impact of job stress on employee job satisfaction a study on telecommunication sector of Pakistan." Journal of Business Studies Quarterly 2, no. 3 (2011): 50.

[21] Ellickson, Mark C., and Kay Logsdon. "Determinants of job satisfaction of municipal government employees." Public Personnel Management 31, no. 3 (2002): 343-358.

[22] Phillips, Jack J., and Adele O. Connell. Managing employee retention: a strategic accountability approach. Routledge, 2003.

[23] García-Bernal, Javier, Ana Gargallo-Castel, Mercedes Marzo-Navarro, and Pilar Rivera-Torres. "Job satisfaction: empirical evidence of gender differences." Women in management review (2005).

[24] Arnold and Feldman (1996) “Organizational Behavior”. Mc Graw Hill

[25] Nwagwu, Williams E. "Job Satisfaction of Information Technology Artisans in Nigeria." Mousaion 36, no. 2 (2018)

[26] Kumar, P. M. Job Satisfaction Among Permanent and Contractual Information Technology Workers. IACIS, (2002):362-366.

[27] Parvin, Mosammod Mahamuda, and MM Nurul Kabir. "Factors affecting employee job satisfaction of pharmaceutical sector." Australian journal of business and management research 1, no. 9 (2011): 113.

[28] Jan, N. Akbar, A. Nirmal Raj, and A. K. Subramani. "Employees' Job Satisfaction in Information Technology Organizations in Chennai City-An Empirical Study." Asian Journal of Research in Social Sciences and Humanities 6, no. 4 (2016): 602-614.

[29] Joshi, Harisha G. "Quality of Work Life Among it Professionals in Sme's in Select Cities of India." GSTF Journal of Law and Social Sciences (JLSS) 1, no. 1 (2012): 151.

[30] Korsakienė, Renata, Asta Stankevičienė, Agnė Šimelytė, and Milda Talačkienė. "Factors driving turnover and retention of information technology professionals." Journal of business economics and management 16, no. 1 (2015): 1-17.

[31] Lin, S. C., & Lin, J. S. J. (2011). Impacts of coworkers relationships on organizational commitment-and intervening effects of job satisfaction. African Journal of Business Management, 5(8), 3396-3409.

[32] Pook, L. A., Füstös, J., & Marian, L. (2003). The impact of gender bias on job satisfaction. Human Systems Management, 22(1), 37–50.

[33] Rast, Sadegh, and Azadeh Tourani. "Evaluation of employees' job satisfaction and role of gender difference: An empirical study at airline industry in Iran." International Journal of Business and Social Science 3, no. 7 (2012).

[34] Ghazzawi, Issam. "Does age matter in job satisfaction? The case of US information technology professionals." Journal of Organizational Culture, Communications and Conflict 15, no. 1 (2011): 25.

[35] Kowal, Jolanta, and Narcyz Roztocki. "Gender and job satisfaction of information technology professionals in Poland." In 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 3625-3634. IEEE, 2016.

[36] Clark AE. Job satisfaction and gender: why are women so happy at work? Labour Econ. 1997;4(4):341–72.

[37] Bordin, Carina, Timothy Bartram, and Gian Casimir. "The antecedents and consequences of psychological empowerment among Singaporean IT employees." Management Research News (2007).

[38] Lim, Sook. "Job satisfaction of information technology workers in academic libraries." Library & Information Science Research 30, no. 2 (2008): 115-121.

[39] Lumley, E. J., Melinde Coetzee, Rebecca Tladinyane, and Nadia Ferreira. "Exploring the job satisfaction and organisational commitment of employees in the information technology environment." Southern African business review 15, no. 1 (2011).

[40] Varma, Aparna J., Kotresh Patil, Ravishankar S. Ulle, A. N. Santosh Kumar, and T. P. Renuka Murthy. "An empirical study on job satisfaction and employee loyalty." Journal of Emerging Technologies and Innovative Research 5, no. 8 (2017): 780-791.

[41] Lee, P. C. B., B. W. H. Chan, and J. C. M. Lee. "Emotional intelligence and information technology professionals." In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 140-144. IEEE, 2017.

[42] Wong, Anthony, and Canon Tong. "Evaluation of organizational commitment models and their components in Asian cities." International Journal of Human Resource Studies 4, no. 2 (2014): 66.

[43] Kumar, Rajesh, Roshan Lal, Yashu Bansal, and Saran K. Sharma. "Technostress in relation to job satisfaction and organisational commitment among IT professionals." International Journal of Scientific and Research Publications 3, no. 12 (2013): 1-3.

[44] Adebiaye, R. "Predictive Factors OF JOb Satisfaction Levels amongst it Professionals in The United States." (2018).

[45] Misra, Sunil, and Kailash BL Srivastava. "Team-building competencies, personal effectiveness and job satisfaction: The mediating effect of transformational leadership and technology." Management and Labour Studies 43, no. 1-2 (2018): 109-122.

[46] Spann, Charlene Stacey. "The Relationships of Role Conflict and Role Ambiguity With Job Satisfaction in Non-Managerial IT Professionals in Matrix Organizations." PhD diss., Grand Canyon University, 2018.

[47] Vennila, M., and Dr K. Vivekanandan. "A Study on How Emotional Dissonance Impact Work Exhaustion, Job Satisfaction and Turnover Intention among It'Professionals." International Journal of Management 8, no. 1.

[48] Ben.Salamah, Fai (2017). "Job Satisfaction Factors of IT Sectors Employees, in Kuwait" [Unpublished Master's thesis]. Kuwait University.

[49] Hu, Li-tze, and Peter M. Bentler. "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives." Structural equation modeling: a multidisciplinary journal 6, no. 1 (1999): 1-55.

Fuzzy based Search in Motion Estimation for Real Time Video Compression

Upendra Kumar Srivastava[1]
Research Scholar, Dept. of CSE
IFTM University, Moradabad
India

Rakesh Kumar Yadav[2]
Assistant professor, Dept. of CSE
IFTM University, Moradabad
India

***Abstract*—Video compression ratio, quality and efficiency are determined by the motion estimation algorithm. Motion estimation is used to perform inter frame prediction in video sequences. The individual frames are divided into blocks the motion estimation is computed by a video codec such as H.264. A video codec computes the displacement of block between the previous frame (reference frame) and the current frame, for each block in current frame the best motion vector is determined in the reference frame as a block belongs to a current frame. In this, research paper, a novel technique has been presented for motion vector calculation, using fuzzy Gaussian membership function. The motion estimation block uses fuzzy membership function to estimate the connectedness of different blocks of the current frame to that of the reference frame The fuzzy decision matching is done based on the matching criterion and the best matching block is selected. The motion vectors are thus calculated with respect to the reference frame. The fuzzification process produces optimally matched blocks, which are then utilized to calculate the motion vectors of the predicted frame. Using fuzzy based search the search area is automatically updated and adaptive search steps provides an optimized result of search. As in real time streaming no file is exchanged during the transmission user is not able to download the file the only way for smooth transmission is frame management fuzzy based search for the motion estimation provides a better compression for the predicted frames.**

Keywords*—Fuzzy logic; motion estimation; compression; current frame; reference frame; predicted frame

I. INTRODUCTION

High resolution images and videos have been made accessible to everyone by the use of the technology. Everyday millions of videos and images are being generated by users over the internet, which is being shared across different platforms freely and fastly. A video consists of an image series known as frames. Image frames need to be stored and transmitted in time and space using broadband transmission. This includes a modern video encoding format to safely and easily share video data in real time. H.261, H.263, H.264, and MPEG1, MPEG2, and MPEG4 are common video codec formats. These standards commonly use temporal redundancy reduction to enable video compression [1].

Motion estimation (ME) in general video-coding systems can effectively eliminate time-redundancy between adjacent videos. ME is also used as an integral part of a video encoder since many computing resources are needed. At the same time, ME contributes in particular to the difficulty of the encoder for its various block sizes and fractional pixel precision motion quest in the video encoding format H.264 [2]. Two main motion vector estimation techniques currently are available pel-recursive algorithm and the block matching algorithm. The pel-recursive technique offers a motion vector estimation method for each pixel that consists of a current frame pixel position in the previous frame [3]. Each picture is divided into a fixed size that does not overlap rectangular blocks of either current or reference frames in the corresponding block algorithms. These blocks are then matched, based on some cost function, to find the best matching block, for which motion vectors are calculated. It is very apparent that only objects displaying motion will change their location in the frames between two frames, while the remaining context remains unchanged. The discrepancy between the present frame and the frame of reference is a residual frame. This contains the details of the frame on which the modifications take place. The encoder describes the model that defines the movement of objects in the system, measuring the forecast frame or motion picture. The video coding standards were primarily used for block matching algorithms for the motion calculation. The hardware is easy to incorporate and predicts movement in real time.

Many strategies exist to find the best matching block. The full search algorithm is the easiest algorithm to find a block with minimal SAD in the reference frame. It's a basic routine that compares the best-matched block to a block in any search area. Although this algorithm seems simple, complicated computations are required which prevent it from being a real-time scheme [4]. Several algorithms for fast search and real time deployment were proposed to resolve the limitations of the complete search algorithm. Some of these algorithms look for the matching block only in a set of blocks in the search area. The logarithmic search, three step searches, four step search, diamond search, and octagon search algorithms are some of these algorithms. In all of these algorithms, unimodal error surface assumption is made. The pixel redundancy within frames of a video sequence may differ depending on the movement in the video. Thus regular boundary conditions to determine the blocks may not as effective. The uncertainty in pixel redundancy can be alleviated by using fuzzy techniques. Several fuzzy-based motion estimation have also been presented by the researchers. Spatio-temporal fuzzy search algorithm using a look-up table structure (LUT) is employed [5].

II. Literature Review

There has been an innumerable number of researches in the field of motion estimation in the last 50 years or so. In this section, some latest research works have been discussed. Yun Cheng et al [6] suggested an algorithm named Modified Diamond Quest. MDS uses Small Diamond Search Pattern (SDSP), which evaluates if the MBD (Minimum Block Distortion) is the original search hub. Where the MBD point is not placed in the search centre, the following search stage would use Simplified Large Diamond Search Pattern (SLDSP). If the MBD point is not within the circular spectrum of a single-pixel radius, SLDSP would be continuously used to find the best equivalent block with a large vector until it becomes the MBD point. Finally, SDSP shall be taken to boost the motion vector, particularly with simple and slow motion vectors for certain video sequences.

A. Anusooya Devi et al. [7] suggested the algorithm entitled 'Efficient Motion Estimation Modified Diamond-Square Search Technique.' This manual includes an updated diamond search algorithm that updates the two DS search patterns. Compared to current search algorithms, the MDSS algorithm is advantageous since the amount of search objects used is reduced while the video quality is maintained. Moreover, relative to the diamond search algorithm, it attempts to accelerate the search.

A Fuzzy Logic Based Three Step Search Algorithm for Motion Vector Estimation [8] was proposed by Suvojit Acharjee and Sheli Sinha Chaudhuri. A fuzzy dependent logic has been introduced into this three step search algorithm. This is a superior algorithm than the Four steps (FSS), the Three step search (TSS) algorithm, the New Tree step search (NTSS).

The Fuzzy Logic Based Four Step Search Algorithm for Motion Vector Estimation. Suvojit Acharjee and Sheli Sinha Chaudhuri [9]. A fuzzy membership value added by strength for each block is used in the Four Step Search algorithm based on fuzzy logic. A value that determines whether the macro block is in the darker or lighter area is determined from the intensity values of the pixels within a macroblock. Only if the macro block's macro frame macro membership value is beyond the permitted macro block region of the current frame will the search continue. The pattern of quest and the other stage is like four stages of the search.

The proposed Fuzzy Thresholding Quick Motive Estimating Scheme for Video Coding was proposed for Fuzzy Thresholding Cheng et al. [10]. The suggested algorithm is an early termination scheme based on fluctuating inference threshold values. Using the MDGDS algorithm search patterns, we used the fluctuating inference variables for the MDGDS three-round alternate search pattern. It is decided before each search round to avoid needless computation that a search is to be terminated at an early stage. In contrast with the MDGDS, the proposed algorithm will reduce the average considerable number of search points. The algorithm increases the motion prediction greatly.

Y. Pattnaik et al. [11] recommended the use of the adjacent blocks to predict the motion vector of the block. They implemented a sorting search algorithm that is more likely to aid in the prediction by using the motion vectors of adjacent lines. With the aid of these motion vectors, a search centre is located and around it, a search window is mounted. The search approach was compared by the authors using a particular neighborhood combination and after a detailed analysis, the sorted algorithm was found to produce the other current PSNR and computing algorithms.

The Fuzzy logic inference system-based hybrid prediction model for the wireless 4k UHD 4k H.265 coded video streaming was proposed by Mohammed Alreshoodi, et al. [12]. The calculation techniques available that follow a complete reference model are inefficient for streaming in real time, as the original video sequences on the recipient side are required. Investigations of service quality (QoS) parameters in the experimental setting for 4kUHD H.265 coded video transmission; secondly, an objective model based on the fuzzy logic inference method is created, with the goal of predicting the visual quality by mapping of the calculated quality of experience parameters with QoS.

For high delay applications of HEVC, Davoud Fani et al. [13] suggested an algorithm for GOP level fuzzy rate management. A Rate Control Algorithm (RCA) has been developed with this algorithm for high-delay HEVC Standard applications with buffering constraints. This RCA is fitted with a fluid controller and a simulated buffer. The fluctuation of the quantization parameter (QP) is designed to eliminate variability when the buffer restriction is complied with. For each pictures category (GOP), it determines a QP basis in order to avoid unwanted variations of the QP at the GOP stage.

The new quick motion evaluation algorithm was developed by Masahiro Hiramori et al. [14]. It concurrently scans 4-pixel groups and uses the value concatenated with the exclusive OR of the low 6-bit absolute upper 2-bit gap. The search accuracy results reveal that, as opposed to the search-related algorithm with a 4-bit absolute difference accumulator, the cumulative difference is improved to 4 of 7 video sequences. The synthesis findings have seen a 61 percent decrease in the required loop, a 15.2 percent decrease in the circuit size, and the operating frequency is improved from 334.67 MHz to 616.90 MHz relative to a total 4-bit absolute.

C. Wu and J. Wu and J. Huang [15] implemented the mobile application motion prediction root predictive pattern search algorithm. The adaptive Root pattern search algorithm is combined to increase the accuracy of the image and reduce search points for two kinds of predictive patterns. The motion vectors of top-left and upper-middle macroblocks are chosen as candidates of ARPS if the block is placed on the right-hand side of the picture. As otherwise, ARPS candidates are the motion vectors of the macro-blocks upper-left, mid-right, and top-right. Where motion vectors are introduced into the algorithm in the previous and neighboring blocks, the trend of the surrounding blocks and the probability for trapping in the local optimum decreases. The results of experiments demonstrated better than other block matching processes, particularly for large and quick motion chips, the image quality of the proposed system.

Arnaudov and Ogunfunmi proposed Fast Motion Prediction adaptive search patterns for HD video. The algorithm tried, not based on the video set, to make the search mode versatile or adaptable for each scene within a given video. As for the writer, it will have a certain output penalty relative to a fixed pattern scan. Every 'I' frame is taught a new pattern. It has emerged from the findings that the Adaptive Search Pattern makes around 10% - 70% of the PSNR difference between current fixed and complete search algorithms [16].

Nijad A-Najdawi [17] suggested a real-time video encoding device that can render immersive, including video conferencing, inexpensive, real-time applications. The proposed algorithm looks at frequency domain motion estimation. Block matching is conducted in the frequency domain, where a group of carefully selected frequencies is checked to accurately classify each block.

Ali Al-Naji et al. [18] suggested quality video measurement index based on the fuzzy inference system, suggesting a new solution based on a floating interface system known as the quality assessment system (QES). As inputs to three fluctuating logic controller systems, their feedback to another fluctuating logic controller system was used as inputs to achieve nine quality metrics; PSNR, visual signal-to-noise ratio; weighted signal-to-noise ratio, structural similarity (SSIM), multi-scale SSIM, uniform image quality index, visual information fidelity; and noise quality analysis (IFQA) Despite the inability of some IQA approaches to provide the quality output of the input video in certain cases, this approach leads to the obtaining of a specific quality index.

Linh Van Ma, et al. [19] suggested an Adaptive Streaming algorithm to boost mobile data efficiency in order to reduce DASH's entropy rate of Bitrate Fluctuation. Dynamic adaptive Hypertext Transmission Protocol (HTTP) streaming is a state-of-the-art video streaming technology that while always and constantly evolving, has one downside. The quality of viewing of videos fluctuates along with changes in the network which could decrease service quality. The average moving bandwidth and buffer values are first determined for a given time. In order to deduce the importance of the video quality representation in the following request, a fuzzy logic method is used based on discrepancies between actual and average values. The entropy speed is often used to calculate the predictable/stabilizing of a bandwidth measurement chain. The experiment leads to decreased video quality variability in contrast with the current approaches and increased 40% of bandwidth consumption.

The suggested modification of the Fuzzy logic-based performance enhancement scheme of DASH (mFDASH) has been proposed by Hyun Jun Kim et al. [20]. By changing the Fuzzy Logic Controller (FLC) for the next line, a more acceptable bandwidth is calculated for the proposed scheme by using the history-based TCP Throughput Calculation, than for FDASH. In addition, mFDASH decreases the number of shifts in the video bit rate by using the SBFM section and uses Launch Function to produce high-quality videos at the very early stage. Finally, the Sleeping Mechanism is used to prevent the predicted overload of buffers. The NS-3 Network Simulator had been used to check mFDASH results. The MFDASH displays a buffer overflow not assured in the FDASH within a restricted buffer capacity. Of the three systems, mFDASH presents DASH consumers with the best quality.

Adaptive Order Cross–Hexagonal Quest for H.264 in motion estimate suggested by Bachu Srinivas and K Manjunathachari [21]. The algorithm uses a smaller cross-shaped model before the first step of a square pattern and in subsequent steps replaces the square pattern with the hexagonal search patterns. The patterns of searches help locate the best matching block, regardless of a large number of search points. The matching points can be measured using the speed and distortion parameters using a fluid-based tangent weighted function. In order to reach visual quality and distortion targets, the suggested approaches are used successfully in the block estimation process.

Srinivas Bachu and N. Ramya Teja have also suggested "Fuzzy Adaptive Selection Mode for H.264 Video Coding based Holoentropy". The main downside, as indicated by the developers, in H.264 is a detailed check over the prediction of the interlayer to obtain the best rate distortion. A new approach for interdiction mode selection, based on the fuzzy holoentropy, has been implemented to reduce the overall measurement due to a comprehensive search on the mode prediction process. In order to determine mode, the device uses pixel values and probabilistic distributions of pixel symbols. This selection of adaptive mode is made possible by the consideration of the pixel values of the current block to be coded using the fuzzy holoentropic for the motion-compensated referential block. The mode judgment that is adaptively chosen will minimize the time of the computation without impacting frame vision [22].

III. Proposed Methodology

Real-time videos involve slow and fast content mixtures of motions. No set quick-block matching algorithm will essentially eliminate the temporal redundancy of wide-motion video sequences. Larger motions warrant a bigger search parameter but make the motion estimation more costly. The complete search motion estimation algorithm coincides with all potential displaced blocks in the reference frame's field of search, among all block matching algorithms, to find a block with minimal distortion. In order to perform a full search, a huge amount of calculation is required. Adaptive step size should also be used to obtain actual motion vectors. The fuzzy logic method can be used to calculate motion by adopting measures. **Fig. 1** shows the block diagram of the proposed model.

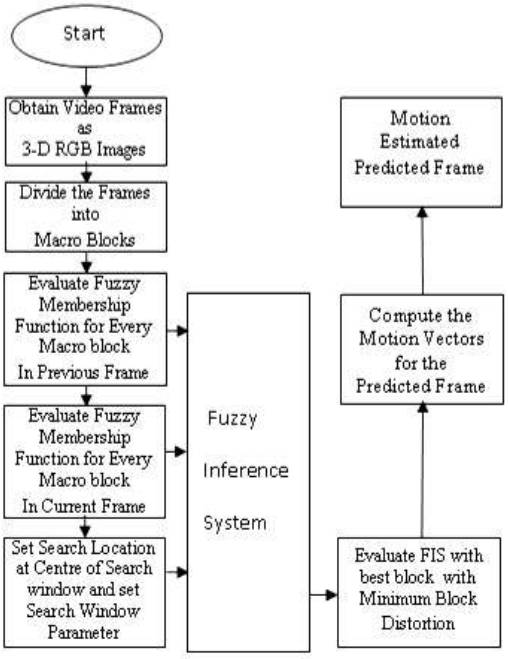


Fig. 1. Block Diagram of Proposed Technique.

The motion estimation based methodology is based on selecting suitable blocks in the video frames. As shown in Fig. 1, the current frames and the reference are both subjected to blocking, i.e. dividing the individual frames into smaller sub-frames or blocks. The motion estimation block uses fuzzy membership function to estimate the connectedness of different blocks of the current frame to that of the reference frame. Gaussian membership function has been used to evaluate the fuzzy membership values for every macroblock. The fuzzy decision matching is done based on the matching criterion and the best matching block is selected. The motion vectors are thus calculated with respect to the reference frame, which can then be used further for facilitating the video compression. The definition of a fuzzy set begins with fuzzy logic. A fuzzy set has no narrow, specifically defined limit. Elements with only a partial membership can be included. A function that defines the extent to which a certain input is part of a set. The membership degree implies that the production is often restricted to a membership function between 0 and 1. Also referred to as a membership or membership category.

IV. Proposed Algorithm

The key factor of the proposed algorithm is adaptive step size search minimize the search cost because it does not have the fixed steps it depends on the fuzzy membership (Gaussian Membership Function)of each pixel and the value of sum of difference. These steps are followed by the algorithm Fig. 2 shows the initial point of the search and Fig. 3 the updated search point and the new search area.

Step 1: This algorithm tends to reduce the search steps for this an adaptive step search strategy is taken and the key factor is the sum of absolute differences (SAD) which is a measurement of the similarities between the blocks which are taken for comparison as block size 8x8 or 16x16 the absolute difference between each pixel in the reference frame block and the corresponding pixel in the block of target frame. Unlike the other algorithm here the SAD is calculated by the Gaussian membership function (GMF) which is assigned for macroblock for previous frame and the macroblocks of target frame for each pixel and a membership data matrix is created.

Step 2: Start searching for the pixel with minimum SAD as compared to target frame the centre point of membership function is decided on the basis of minimum SAD and a search area is constructed and with a updated centre of the membership function a new search area is constructed.

Step 3: The Sum of Absolute Differences (SAD) parameter is utilized to obtain the motion vectors of the moving blocks. The blocks for which minimum SAD is obtained constitute the candidates for motion vectors.

This work determines the membership values of block coefficients to show the consistency of the coefficients by using numeric values to determine the fluctuating and unsure pixel quality. Fuzzy sets can then be employed to measure the degree of value for any pixel.

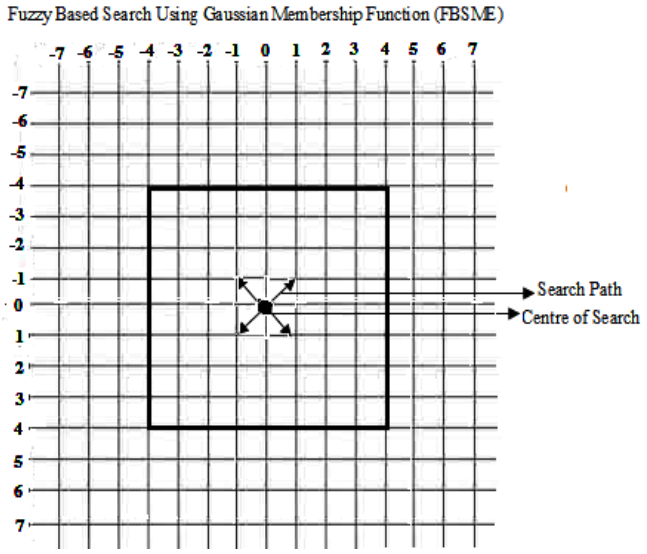


Fig. 2. Sketch Diagram of Proposed Algorithm Initial Search.

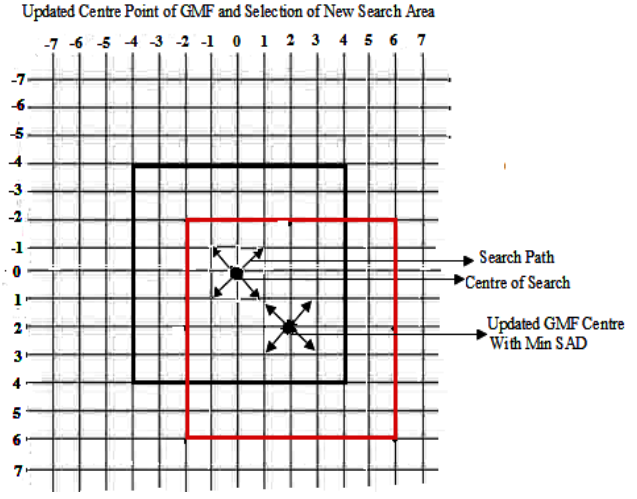


Fig. 3. Sketch Diagram of Proposed Algorithm Updated Search.

The Gaussian membership function has only two parameters that can be determined by macro block pixels; thus, image pixels distribution can be treated as a regular distribution according to the membership function. A macro block with size "M x N" can be treated as the data matrix, and for this the corresponding membership matrix can be obtained using Gaussian membership functions. This membership matrix contains an array of fuzzy sets, namely the fuzzy set of corresponding pixels values. $u(im_{ij})$ represents the degree of membership each pixels [23], as in eq 1. After formulating the membership function, each crisp pixel value im(i,j) is assigned as a membership value $u(im_{ij})$ value which is the corresponding membership degree of the fuzzy set.

$$u(im_{ij}) = e^{\frac{-(imij-c)^2}{2\varphi\sigma^2}} \quad (1)$$

Where imij represents the macroblocks' intensity, φ is the amplification factor, σ the macroblocks standard deviation and their width is the GMF, c the centre of the GMF and the macroblock's average value is described.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(im(i,j)-c)^2}{MxN}} \quad (2)$$

$$c = \frac{1}{MxN}\sum_{i=1}^{M}\sum_{j=1}^{N}(im(i,j) - c) \quad (3)$$

Where im(i,j) is the pixel value at position (i, j),M and N is the size of the block.

The membership functions as defined in above equations are calculated for both current frame and the previous or reference frame. The fuzzy decision for the predicted frame motion vectors is calculated by finding difference of the two membership values.

em=(abs(ur(imij)-uc(imij))) (4)

where ur(imij) is membership of reference block and uc(imij) is membership of current frame blocks. The sum of absolute difference of membership value of all blocks is the de-fuzzification expression.

$$SAD = \sum_{k=1}^{B} em \quad (5)$$

where "B" represents all the macroblocks. The Sum of Absolute Differences (SAD) parameter is utilized to obtain the motion vectors of the moving blocks. The blocks for which minimum SAD is obtained constitute the candidates for motion vectors.

V. Result and Discussion

The proposed algorithm was implemented using MATLAB software and tested with 'football.mp4' video sequence**.** The frames have been derived from the original video sequence which is of the size 352x288. The bit rate of the video is 4Mb/s. In most of the researches done earlier a grayscale or monochrome version has been chosen for analysis but here in this work, colored frame retrieved as from the original video has been utilized for the analysis. Fig. 4, 5 and 6 show comparative results of Full Search(FS), H.264, Three Step Search(3SS) and proposed Algorithm using different search area "p", and block size "b". **Fig.** 4(a), 4(b), 4(c), 4(d), 4(e), 4(f), 4(g), 4(h), 4(i), 4(j), 4(k) and 4(l) show the predicted frame, residual frame and motion vector plot for FS,H.264, 3SS and the Proposed Method for p=8 and b=8. Similarly in **Fig. 5**(a), 5(b), 5(c), 5(d), 5(e), 5(f), 5(g), 5(h), 5(i), 5(j), 5(k) and 5(l) similar results for the three techniques and a proposed fuzzy method have been shown for, p=8 and b=16 and in **Fig.** 6(a), 6(b), 6(c), 6(d), 6(e), 6(f), 6(g), 6(h), 6(i), 6(j), 6(k) and 6(l) all these three technique and a proposed fuzzy method have been applied for the p=16 and b=16.

A. Results for clip football.mp4 sequence Search area p=8, Block Size, b=8.

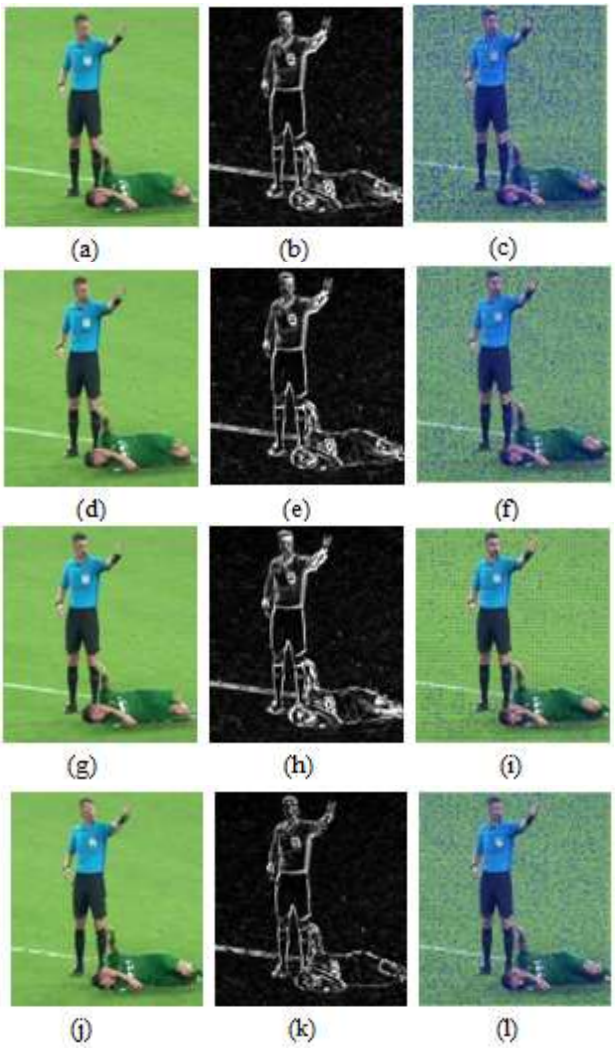


Fig. 4. (a) Predicted Frame(FS), (b) Residual(FS), (c) Motion Vector(FS), (d) Predicted Frame(H.264), (e) Residual(H.264). (f) Motion Vector(H.264), (g) Predicted Frame3SS), (h) Residual(3SS), (i) Motion Vector(3SS), (j) Predicted Frame(Proposed) (k) Residual(Proposed) (l) Motion Vector(Proposed).

B. Results for clip Football.mp4 Sequence Search Area p=8, Block Size, b=16.

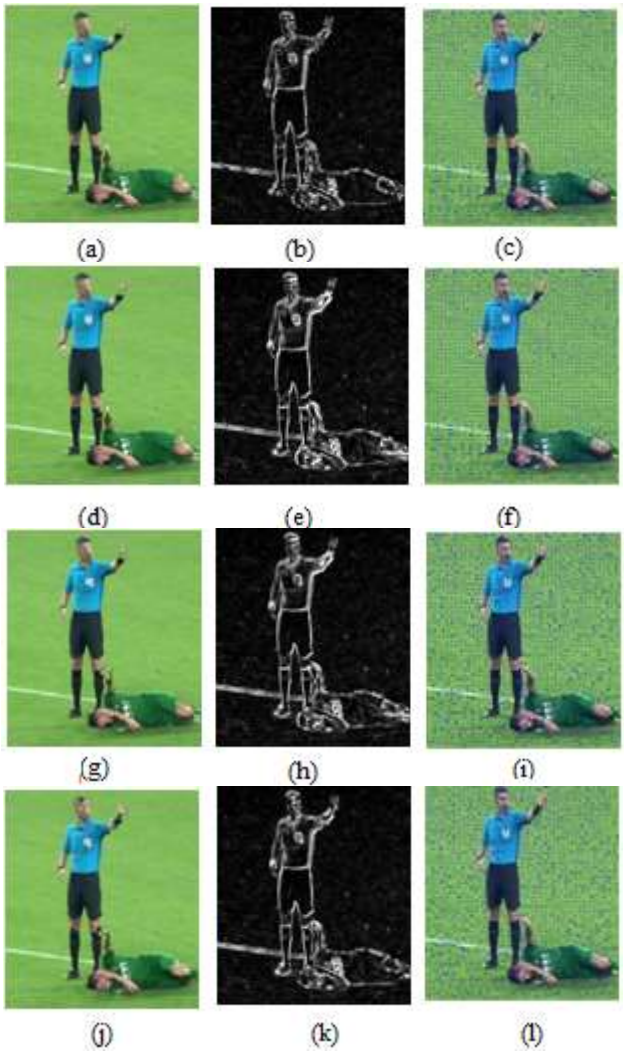
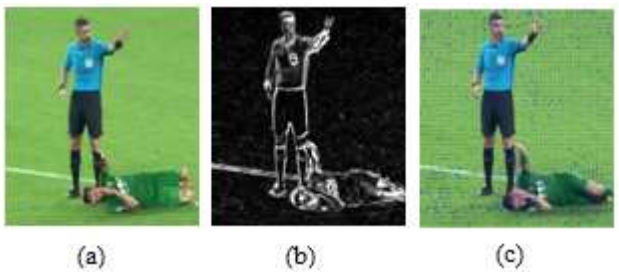


Fig. 5. (a) Predicted Frame(FS), (b) Residual(FS), (c) Motion Vector(FS), (d) Predicted Frame(H.264), (e) Residual(H.264), (f) Motion Vector(H.264), (g) Predicted Frame3SS), (h) Residual(3SS), (i) Motion Vector(3SS), (j) Predicted Frame(Proposed), (k) Residual(Proposed), (l) Motion Vector(Proposed).

C. Results for clip football.mp4 sequence Search area p=16, Block Size, b=16



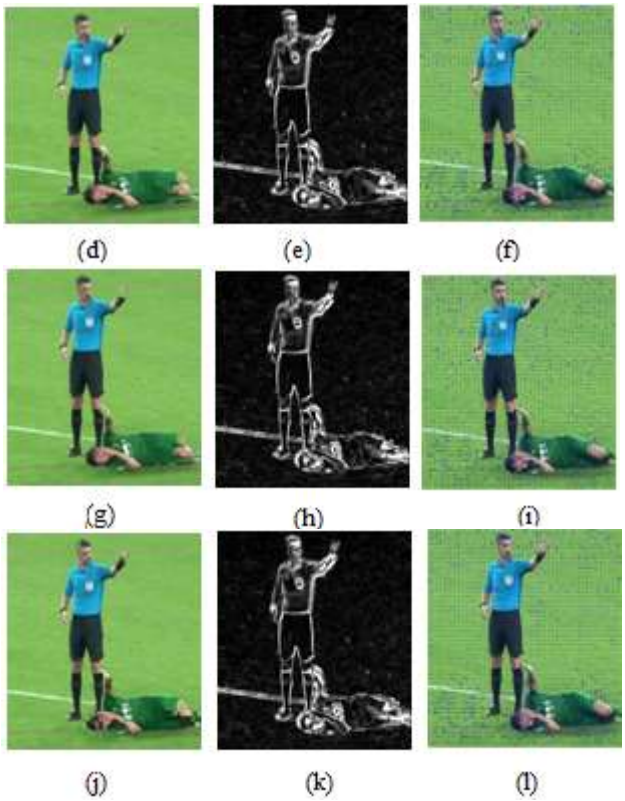


Fig. 6. (a) Predicted Frame(H.264), (b) Residual(H.264), (c) Motion Vector(H.264), (d) Predicted Frame(H.264), (e) Residual(H.264), (f) Motion Vector(H.264), (g) Predicted Frame3SS), (h) Residual(3SS), (i) Motion Vector(3SS), (j) Predicted Frame(Proposed), (k) Residual(Proposed), (l) Motion Vector (Proposed).

D. PSNR Analysis Comparison Table Football.mp4

The Peak Signal-to-Noise Ratio is expressed for the ratio between two values the maximum possible value of a signal and power of distortion the higher the value the image quality is better because of the Mean square error between the original frames and predicted is very low. Here the PSNR value is computed between the target frame and predicted frame .While computing the PSNR value of different algorithm as Full search, three step search and H.264 the proposed algorithm shows that it has higher PSNR value than other algorithms. The comparison results are shown in Table I and the related graph in **Fig. 7**.

TABLE I. PSNR COMPARISON TABLE (FOOTBALL.MP4)

Method	**p=8,b=8**	**p=8,b=16**	**p=16,b=16**
FS	69.4792	68.7755	68.8051
H.264	67.0061	66.8010	66.7552
3SS	69.8856	69.0901	69.0413
Proposed Fuzzy	70.6926	69.4762	69.3484

E. PSNR Graph Representation Football.mp4

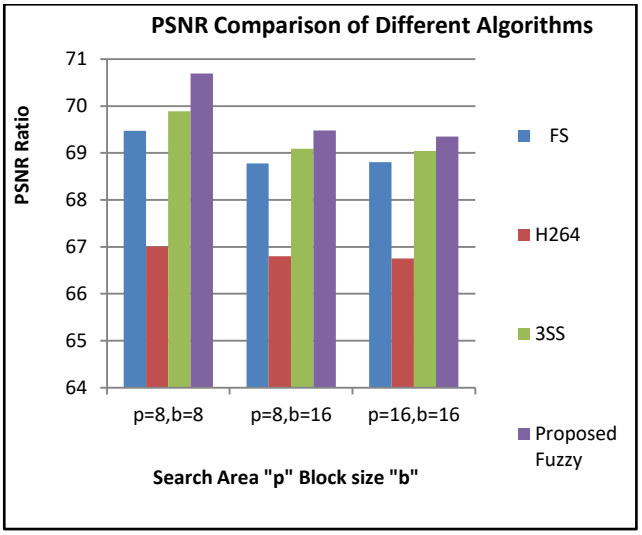


Fig. 7. PSNR Comparison Graph (Football.mp4).

F. SSIM Comparison (Football.mp4)

Structural Similarity index which measures the difference between predicted frame and target frame, based on visible structures in the image and perceptual metric is given in the Table II and the related graph in Fig. 8.

TABLE II. SSIM COMPARISON TABLE (FOOTBALL.MP4)

Method	p=8,b=8	p=8,b=16	p=16,b=16
FS	0.9996	0.9994	0.9994
H264	0.9989	0.9988	0.9988
3SS	0.9996	0.9995	0.9995
Proposed Fuzzy	0.9998	0.9996	0.9996

G. SSIM Comparison Graph (Football.mp4)

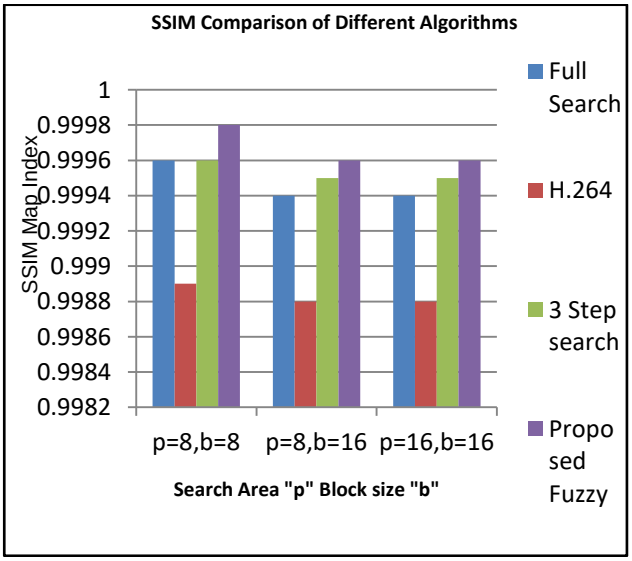


Fig. 8. SSIM Comparison Graph (Football.mp4).

VI. CONCLUSION

In Table I, the PSNR value for the different algorithms has been evaluated for various values of p & b parameters. For p=8,b=8, the proposed fuzzy based method achieves a PSNR of 70.6926 as compared to 69.4792,67.0061,69.8856 of full search, H.264 and three step search algorithms for p=8,b=16 and p =16,b=16, the proposed fuzzy based search algorithm achieves better PSNR values of 69.4762 and 69.3484 respectively. It proves that the fuzzy based search provides an optimal search and maintains the frame quality which is more suitable in real time video streaming. Thus it can be asserted that the proposed algorithm achieves a better PSNR as compared to other algorithms. In the same way SSIM is also found better than other algorithms. This proves the fact that the proposed motion estimation algorithm is more suitable for the compression standards to yield the optimized performance.

ACKNOWLEDGMENT

I express my sincere gratitude towards Assistant Prof. Dr. Rakesh Kumar Yadav, Department of Computer Science and Engineering IFTM University, for his valuable suggestions and thanks to Head of Department of Computer Science and Engineering IFTM University, Moradabad for gave us necessary facilities for the implementation of this research work.

REFERENCES

[1] Naseer Al-Jawad, Johan Ehlers, & Sabah Jassim, "An efficient real-time video compression algorithm with high feature preserving capability" EC SecurePhone project IST-2002-506883, 2002.

[2] T. Fryza, "Introduction to Implementation of Real Time Video Compression Method", In the Proceedings of the IEEE Conferences 2008 15th International Conference on Systems, Signals and Image Processing , Bratislava, Slovakia, pp. 217 - 220, 2008.

[3] Suvojit Acharjee ,Nilanjan Dey and Debalina Biswas et al., "An efficient motion estimation Algorithm using division mechanism of low and high motion zone ", In the Proceedings of the IEEE Conferences 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) , Kottayam, India, pp. 169 - 172, 2013.

[4] Yun-Gu Lee, "Early search termination for fast motion Estimation ", EURASIP Journal on Image and Video Processing , Vol. 2015 ,Article 29. pp. 1 –10, 2015.

[5] S.M.R. Soroushmehr, S. Samavi and M. Saraee., "Fuzzy block matching motion estimation for video compression ", In the Proceedings of the IEEE Conferences 2009 IEEE 9th Malaysia International Conference on Communications (MICC) , Kuala Lumpur, Malaysia, pp. 1 - 4, 2009.

[6] Yun Cheng and Min Wu in ,"A Fast Motion Estimation Algorithm Based on Diamond and Line/Triangle Search", In the Proceedings of the IEEE Third International Conference on Pervasive Computing and Applications, Alexandria, Egypt ,pp. 537-542 2008.

[7] A.Anusooya Devi , M.R.Sumalatha, N.Mohana Priya, B.Sukruthi and M.Minisha "Modified Diamond-Square Search Technique for Efficient Motion Estimation", In the Proceedings of the IEEE Conferences International Conference on Recent Trends in Information Technology (ICRTIT) Chennai, Tamil Nadu, India pp. 119-1153 ,2011.

[8] Suvojit Acharjee,Sheli Sinha Chaudhuri , "Fuzzy Logic Based Four Step Search Algorithm for Motion Vector Estimation", I.J. Image, Graphics and Signal Processing , Vol. 2, pp. 37 – 43, 2012.

[9] Suvojit Acharjee,Sheli Sinha Chaudhuri , " Fuzzy Logic Based Four Step Search Algorithm for Motion Vector Estimation", I.J. Image, Graphics and Signal Processing , Vol. 4, pp. 49 – 55, 2012.

[10] Hung-Ming Chen, Jhong-Kai Lin, Po Hung Chen, Ting-Jhao Jheng , "A Fuzzy Thresholding Early Termination Scheme of Fast Motion

Estimation for Video Coding",IEEE 17th International Symposium on Consumer Electronics , pp. 243-244 , 2013.

[11] Yogananda Patnaik, Dinesh Kumar Singh and Dipti Patra., "A new search method for block motion estimation in Video Compression ", In the Proceedings of the IEEE Conferences 2014 Annual IEEE India Conference (INDICON) , Pune, India, pp. 1 - 5, 2014.

[12] Mohammed Alreshoodi , Anthony Olufemi Adeyemi-Ejeye, John Woods, Stuart D. Walker," Fuzzy logic inference system-based hybrid quality prediction model for wireless 4kUHD H.265-coded video streaming*"* The institution of Engineering and Technology , Vol.4, Issue 6, pp. 296-303, 2015.

[13] Davoud Fani Mehdi Rezaei ,"A GOP-level fuzzy rate control algorithm for high-delay applications of HEVC*"* Signal of Image and Video Processing , Springer Signal. Image Video Process. (SIViP) , pp.1-9, 2016.

[14] Masahiro Hiramori , Ryota Bandou, et al. "A study on Fast Motion Algorithm" In the Proceedings of the IEEE Conferences 2016 IEEE 5th Global Conference on Consumer Electronics , Kyoto, Japan, pp. 1 - 3, 2016.

[15] Chia-Ming Wu and Jen-Yi Huang "Predictive Root Pattern Search Algorithm for Motion Estimation on Mobile Devices " In the Proceedings of the IEEE Conferences 2016 IEEE 5th Global Conference on Consumer Electronics , Kyoto, Japan, pp. 1 - 2, 2016.

[16] Pavel Arnaudov and Tokunbo Ogunfunmi "Adaptive Search Pattern for Fast Motion Estimation in HD Video" 2017 51st Asilomar Conference on Signals, Systems, and Computers , Pacific Grove, CA, USA pp. 1 - 5, 2017.

[17] Nijad A-Najdawi, "Fast Block Matching Criterion for Real-Time VideoCommunication", In the Proceedings of the IEEE Conferences 2017 International Conference on New Trends in Computing Sciences (ICTCS) , Amman, Jordan, pp. 327 - 332, 2017.

[18] Ali Al-Naji, Sang-Heon Lee and Javaan Chahln *"*Quality index evaluation of videos based on fuzzy interface system ", IET Image Processing, Vol. 11, pp. 292-300, 2017.

[19] Linh Van Ma , Jaehyung Park , Jiseung Nam , Hoyong Ryu and Jinsul Kim " A Fuzzy-Based Adaptive Streaming Algorithm for Reducing Entropy Rate of DASH Bitrate Fluctuation to Improve Mobile Quality of Service ", Entropy , pp. 1-18 2017.

[20] Hyun Jun Kim, Ye Seul Son, and Joon Tae Kim " A Modification of the Fuzzy Logic Based DASH AdaptationScheme for Performance Improvement " Wireless Communications and Mobile Computing Vol. 2018, pp. 1-16, 2018.

[21] Bachu srinivas, and k manjunathachari ," Adaptive Order Cross–Square–Hexagonal search and Fuzzy Tangential -weighted Trade-off for H.264 in Motion Estimation", Sadhana , Indian Academy of Sciences ,pp.1-16 , 2018.

[22] Bachu srinivas, and N. Ramya Teja , " Fuzzy Holoentropy-Based Adaptive Inter Prediction Mode Selection for H.264 Video Coding*"*, International Journal of Mobile Computing and Multimedia Communications Vol. 10 No. 2 ,pp.42-60 , 2019.

[23] Qian Jiang, Xin Jin, et al. *"*A Novel Multi-Focus Image Fusion Method Based on Stationary Wavelet Transform and Local Features of Fuzzy Sets*"*, IEEE Access multidisciplinary, open access journal of the IEEE.Vol. 5 ,pp.20286-20302 , 2017.

Using Blockchain based Authentication Solution for the Remote Surgery in Tactile Internet

Tarik HIDAR[1], Anas ABOU EL KALAM[2], Siham BENHADOU[3], Oussama MOUNNAN[4]
Hassan II university, LISER IPI-LRI ENSEM, Paris-France, Casablanca-Morocco[1]
Cadi Ayyad University, ENSA-Team Laboratory, Marrakesh, Morocco[2]
Hassan II University, LISER IPI-LRI ENSEM, Casablanca, Morocco[3]
Ibn Zohr University, FSA-LabSI Laboratory, Agadir,Morocco[4]

***Abstract*—Since the Tactile Internet has been considered as a new era of Internet, delivering real-time interactive systems as well as ultra-reliable and ultra-responsive network connectivity, tremendous efforts have been made to ensure authentication between communication's parties to secure remote surgery. Since this human to machine interaction like remote surgery is critical and the communication between the surgeon and the tactile actor i.e. robot arms should be fully protected during the surgical procedure, a fully secure mutual user authentication scheme should be used in order to establish a secure session among the communicating parties. The existing methods usually require a server to ensure the authentication among the communicating parties, which makes the system vulnerable to single of point failure and not fit the design of such critical distributed environment i.e. tactile internet. To address these issues, we propose a new decentralized blockchain based authentication solution for tactile internet. In our proposed solution, there is no need for a trusted party; moreover, the decentralized nature of our proposed solution makes the authentication immutable, efficient, secure, and low latency requirement. The implementation of our proposed solution is deployed on Ethereum official test network Ropsten. The experimental results show that our solution is efficient, highly secured, and flexible.**

***Keywords*—Tactile internet; blockchain; human to machine interaction; authentication; remote surgery**

I. INTRODUCTION

Emerging industrial trends suggests that the new generation systems would increase the penetration robotics hardware, virtualization technologies and mobile platforms systems. Those new technologies will absolutely switch the role of machines to the human to machine interactions. Especially, the use of the next generation network such as Tactile Internet by mixing ultra-low latency, availability, reliability with high level of security, will represent a revolutionary level of development for society, health, economics and culture.

The mobile internet allowed us to exchange data and make human-to-human relationship. The next step is the Internet of things IoT, which is the interconnection and communication between objects through the internet like sensors and actuators. The Tactile Internet is the next evolution that will not only enable the control of the IoT in real time and low latency. But it will also add a new dimension to human-to-machine interaction by enabling tactile and haptic sensations. In other words, the Tactile Internet is the democratizing of skills and expertise to promote equity between people independently of age, gender and religion.

Now, we summarize some of the key requirements and challenges of industrial Tactile Internet architecture:

Latency: Latency is a measure of delay, in networks domains; it defines the time that takes a data packet to reach its destination. Typically, it is measured as the time taken by data to be transmitted to the recipient and returned to the transmitter. In other words, latency is the time required to make a round trip between two entities. It relies on four delays such as transmission and propagation delay, device-processing delay and storage delay. Ideally, it should therefore be as close to zero as possible. In Tactile Internet environment, latency must not exceed 1 ms in order to ensure real time interaction. Otherwise, the human to machine relationship will be established [1].

Reliability: Tactile Internet actors like, robots and 5G configured smartphone, need a reliable ubiquitous connectivity under all environments such as Tactile Internet environment guarantying high availability, almost 99.999 %, and decreasing the mean time between failures MTBF for optimal operation of all aspect of the applications including maintenance, assembly, construction and repair [1] [2].

Resilience: Several applications like e-Health require the scalability in terms of Tactile Internet architecture, that least should be resilient when several autonomous robot hardware, sensors, and Tactile Internet actors connect through the networks, offering a plethora of recovering from failures and other services. In order to realize that, Tactile Internet networking resilience have to be improved by enabling ubiquitous uptime and fastest main time to repair MTTR, in turn of creating an 'always on' system [3].

Security: The Tactile Internet architecture cannot be secured with the traditional technics of internet technologies. For example, Tactile Internet actors are vulnerable and not secure against the distributed denial of service (DDoS) attacks, which decrease the availability, remote hijacking, cloning attacks and man in the middle. Any single Tactile Internet actor could represent a single point of failure (SPOF) for the entire network and thus damage the availability of data, confidentiality and integrity. Which could cause many disasters especially in health field.

Nobody can deny that Tactile Internet will allow doctors in devastated areas, far from the border, to operate remotely their patients. Therefore, doctors from large hospitals will be able to help colleagues from smaller institutions. That kind of surgery deals with the life and death situations of patients.

In order to make the remote surgery in Tactile Internet environment commercially successful, some factors like security decide the performance of such next generation technologies [25].

Consequently, we have to conceive a model for authentication in order to secure human to machine interactions, like remote surgery, in Tactile Internet environment. Thus, a surgeon can now authenticate to a robot arm using good-shared session key and build a high level of security in communication.

The reminder of this paper is organized as follows: We present firstly an overview of the different related work in Section II, we derive with a detailed description of all components of our architecture including the functions and events in Section III, after we proceed our contribution with an implementation of that solution in Section IV. We present then the security analysis and evaluation of the proposed blockchain based authentication solution in Section V. Finally, we conclude our paper with future work and conclusion.

II. Related Work

Before presenting the related works, we introduce this section with a brief paragraph describing how internet of skills work.

The tactile internet, principally, gives human senses the opportunity to enhance, enable and improve interactions with new technologies.

Haptic interactions will be enabled by the internet of skills using visual feedback. This least will not only include robotic systems and actuating robots that can be controlled in real time. But also, it encompasses the audiovisual interaction.

Different technologies will be mixed by the Tactile Internet, at the network and application level of the open system interconnection (OSI) model. At the edges, robots or 5G configured smartphones will enable the Tactile Internet. Touch in terms of data will be transmitted over a 5G network via the air interface and optic fiber between the e-nodes, while artificial intelligence, especially reinforcement learning, will be enabled close to the user equipment through mobile edge computing (MEC). At the application level, automation, robotics, remote surgery, telepresence, augmented reality (AR) and virtual reality (VR) will all play a part.

Many use cases of authentication schemes have been proposed in different domains. We can consequently use the features of these mechanisms in order to propose our new authentication model for securing one of the most critical use case of the Tactile Internet, which is the remote surgery, which will be the object of section II of our paper.

Challa et al. [4] proposed in their work a user authentication scheme for next generation network like Internet of things applications. The scheme proposed is based on Elliptic curve cryptography (ECC) signature, it also provides user intractability and anonymity features. But, this model requires more computation in internet of things environment.

Hsieh and Leu [5] presented a new model to enhance authentication in the proposed solutions cited in [6] [8]. Wu et al. [8] proposed after a security analysis for Hsieh et al.'s model and gave then a proof that their solution was vulnerable to different attacks like user forgery, offline guessing, physical capture and privileged insider. Furthermore, the proposed model by Leu et al. lacks mutual authentication and does not ensure session key security. To resolve this problem, Wu et al. and Vaidya et al. conceived a user authentication model in the wireless sensor networks WSNs and adopted it to the Internet of Things environment [7] [8].

Li et al. [9] studied the model of Jian et al. [10] and proved that their proposed mechanism was susceptible to key session attack. Then, they improve the solution by proposing an enhanced model for user authentication. Later, He et al. [11] Proposed an architecture based on hierarchical cryptography for the Mobile Healthcare Social Networks. However, owing to the identity based cryptography technic, this solution requires more computation and communication. Another user authentication model is also proposed by Feng et al. [12], their work entitled ideal lattice based anonymous authentication protocol for mobile devices provided a high level of security. Nevertheless, it also need some computation efforts.

Farash et al. [13] provided a scheme for key agreement and user authentication in heterogeneous wireless sensor networks architecture and in Internet of Thing environment. Later, Amin et al. [14] improved performances of the Farash et al. mechanism. After, they gave a proof that their solution knew many security failures. As cited in their work, the Farash et al suffers from offline guessing, specific temporary information leakage and spoofing attacks. Srinivas et al. [15] observed and analyzed in other work that the mechanism of Amine et al. was susceptible to user impersonation, spoofing and stolen smart card attacks. Consequently, they provided an enhanced mechanism for user authentication in WSNs and Internet of Things field in future research [17-18].

Khalil et al. [16] proposed the integration of WSNs into the IoT. Like the WSNs. In the other hand, Yeh et al. [19] proposed Elliptic Curve Cryptography (ECC) technic for user authentication in wireless sensor networks. However, their technic lacks mutual authentication. To overcome this limitation, Shi and Gong [20] Discussed about another elliptic curve cryptography base user authentication model, which will be applied in the wireless sensor networks. Later on, Turkanovic et al. [21] presented a new method to enhance key agreement and user authentication in WSNs. But, their method suffers from many problems such as offline password and identity of guessing, impersonation and smart card stolen attacks. Furthermore, it does not provide secure mutual authentication [22].

Hidar et al. [1], which is our previous work, proposed physical unclonable function to ensure performances and security in the tactile internet; they proposed a mutual authentication protocol basing on the PUFs to resolve the

problem of security with guarantying the same latency, but this solution suffer from the problem of single point of failure (SPOF).

Furthermore, Friedrich Pauls et al. [23] propose a latency-optimized accelerator for hash-based digital signature processing for the Extended Merkle signature scheme XMSS algorithm. Their architecture improves the latency of establishing sessions and the verification into the sub-millisecond range. But it also needs more computational efforts.

Most of the available mechanism presented in this section for user authentication and key agreement are not protected against different attacks. In addition, some of the schemes discussed above are not lightweight, as they require more computational efforts. To summary, the existing related work presented in this section cannot be adequate for ensuring security, especially in a critical case like a remote surgery. Therefore, we need to design a concept for user authentication in the Tactile Internet environment. To the best of our knowledge, we propose a generalized authentication mechanism basing on Blockchain and Smart Contract for remote surgery in a human to machine relationship.

III. Blockchain Mutual Authentication Solution

In this section, we begin with a general background, and then we present our proposed authentication solution for the remote surgery.

A. Blockchain

According to Nakamoto Satoshi in 2008 [24], the blockchain is a distributed database for transactions between entities. All those transactions are stored into ledgers, which ensure security. The non-trusting entities can thus exchange data with each other with a cryptographically verifiable way.

The blockchain paradigm is based on four fundamental blocks:

Source and destination's identifying: All users in a blockchain send and receive transactions with digital identities called addresses. The address must not only give any idea about its owner (Anonymous), but also it should be independent of any given authority (self-generated).

Smart contract: An entity control the condition of auto processing for transactions. In other words, Smart contract.

Transaction: It refers to the act of transmitting data from source to destination. It is generated by sender and broadcasted within the network. All nodes must mine transactions in order to be valid.

Consensus: In blockchain technology, each user or node has absolutely the same ledger as all other users in the network. Consequently, a complete consensus from all nodes is ensured.

B. Proposed Solution

In this part, we provide a description of a generalized user authentication mechanism in the Tactile Internet environment such as remote surgery using our decentralized blockchain user authentication solution.

Our architecture will be described as: a remote surgeon performs a remote surgery on a patient residing in other country, using a robot arm. The surgeon thus does not only need to be physically near to the patient that he operates, but also that least does not need to change his place in critical cases.

To ensure Security and implement our contribution, we propose the following transactions between actors as displaying in Fig. 1:

1) The Surgeon with his remote surgery system authenticates to the smart contract using his wallet address.

2) If the transaction is valid, the smart contract spread a Token access and the IP address of sender. Then, the surgeon and Robot Arm, residing in a smart home, receive the broadcasted data from the blockchain.

3) The surgeon creates a message containing Access token, IP address and the blockchain public key. This package will be signed using the blockchain private key then sent with its corresponding public key.

4) When the robot receives the message, it checks if:

- Both received public keys are similar.
- The signed message is real.
- The public key belongs to the sender address.
- Access token is massively valid.
- The two IP addresses of message and sender are similar.

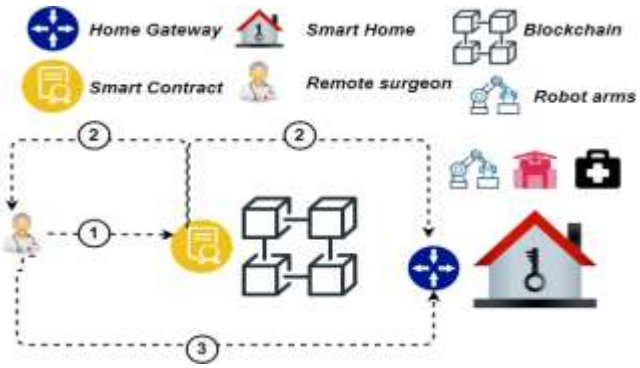


Fig. 1. Steps for user Authentication Model in Tactile Internet based Remote Surgery.

IV. Implementation

We consider an organization that would like to manage the remote surgery using blockchain. First, we create our surgery smart contract using the high-level language programming solidity of Ethereum blockchain [26]. Then, we compile our remote surgery smart contract into Ethereum Virtual Machine (EVM) byte code. Afterwards, we deploy our remote surgery smart contract to the private blockchain (i.e., Ganache [27]) and to the public blockchain (i.e., Ethereum official test network Ropsten [28]). First, we generate a keypair of Externally Owned Account (EOA), the public key (i.e., o.EOA) and corresponding private key (i.e., o.EPK). This keypair of public and private keys are used to create our

remote surgery smart contract (surgery -SC) and execute the functions of the surgery -SC (see Fig. 2). Then, we add via, our smart contract, the authorized users (i.e., remote surgeon) into the smart contract. It includes the smart contract's address and some other information (e.g., surgeon notes). Our surgery smart contract allows to easily add users as well add access policies to the system and to manage and modify the remote surgery access control in a fully decentralized, secure, and transparent manner. Our surgery smart contract ensures the flexibility in the process of adding and removing access control policies.

Moreover, each access control transaction is verified by all nodes of blockchain (i.e., miners), thus ensuring the decentrazlid and trustworthiness of our proposed access control.

To show that our proposed is cost effective, we have estimated the cost of our surgery smart contract as well as the execution of each of its functions. When conducted the experiment, the gasPrice is set to 1Gwei, where 1Gwei=10^{-9} ether, and 1 ether is equal to 379.26 USD. Fig. 3 shows the surgery smart contract functions costs of different functions implement by our proposed access control. We have varied the number of users from 1 to 100. The cost of the execution of different functions of our proposed solution is 0.017, 0.0050.007 And 0.006 USD for add policy, remove policy, delete policy, and check policy functions, respectively. We observe that all the operations have low costs.

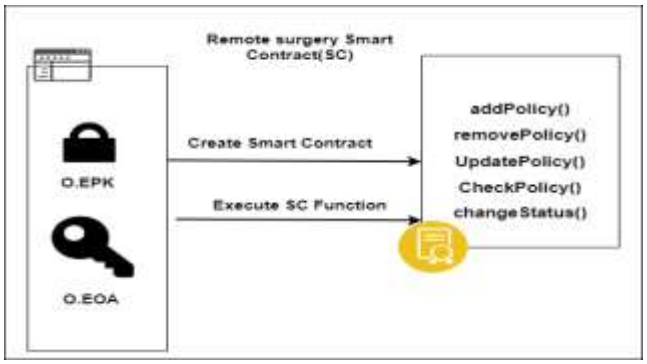


Fig. 2. Surgery Smart Contract.

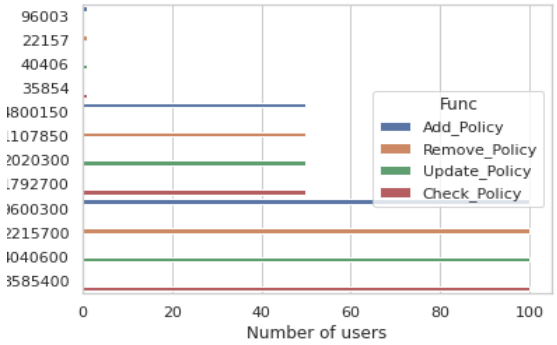


Fig. 3. Surgery Smart Contract Functions Costs.

V. Evaluation and Security Analysis

After the implementation of our proposed solution for ensuring mutual authentication by using blockchain in a Tactile Internet environment. We move to the next section of our paper wish is the evaluation and the security analysis.

In this section, we present an evaluation of our proposed solution in order to assure its quality, we compare it to the previous solutions presented in related work. The metric of this evaluation will be based on if our offered authentication solution solved problems occurring in the other authentication mechanisms proposed in different.

Table I shows a comparison between our solution and the other proposed solutions in section II based on:

- Availability.
- Decentralization.
- Scalability.
- Session key leakage.
- Online and offline password guessing.
- Man in the middle.
- Denial of services.
- Physical capturing of devices.

Table I shows how our proposed solution for the remote surgery is secured againt all the attacks proposed.

TABLE I. Comparison of Security Features and Attacks

Features	Challa	Amin	Li	Jiang	Turkanovic	Farash	Hidar	Our solution
	et al. [4]	et al.[14]	et al.[9]	et al.[10]	et al. [21]	et al.[13]	et al.[1]	
Availability	×	✓	×	×	×	×	✓	✓
Scalability	✓	✓	×	✓	×	×	×	✓
Decentralization	×	×	✓	✓	×	×	×	✓
Privileged insider	✓	×	✓	✓	×	×	✓	✓
Session key agreement	✓	✓	✓	✓	×	✓	✓	✓
Password guessing	✓	×	✓	✓	×	✓	✓	✓
Man in the middle	✓	✓	✓	✓	✓	✓	✓	✓
Denial of service	✓	✓	✓	✓	✓	×	✓	✓
Physical capturing	✓	✓	✓	×	×	×	✓	✓

VI. Conclusion

In this paper, we present a real contribution for Tactile Internet security, which is based on the deployment of the blockchain and smart contract for user authentication in a remote surgery within a Tactile Internet environment. After discussing the existing related work, concerning Tactile Internet and other fields such as wireless secure networks and Internet of things, we showed some vulnerabilities of these proposed solutions, we then propose a general user authentication model by describing highly all various steps needed by remote surgeon in order to get access to the Tactile Internet environment and then operate the patient via the arm robot. After we presented an implementation of our contribution, then we evaluated our work with presenting a comparison between our solution and the others described in related works. In our future work, as the Tactile Internet based remote surgery architecture requires real time reaction, extra low latency and ultra-fast authentication. We are now working on an extension of this paper; we focus on the latency aspect of our mechanism by integrating some technologies like Fog Computing [25] in order to ensure a high level of quality of experience.

References

[1] Hidar, T., El Kalam, A. A., & Benhadou, S. (2019, April). Ensuring the Security and Performances in Tactile Internet using Physical Unclonable Functions. In 2019 4th World Conference on Complex Systems (WCCS) (pp. 1-6). IEEE.

[2] AMAN, Muhammad Naveed, CHUA, Kee Chaing, SIKDAR, Biplab. Mutual authentication in IoT systems using physical unclonable functions. IEEE Internet of Things Journal, 2017, vol. 4, no 5, p.1327-1340.

[3] Fettweis, Gerhard P. "The tactile internet: Applications and challenges." IEEE Vehicular Technology Magazine 9.1 (2014): 64-70.

[4] CHALLA, Sravani, WAZID, Mohammad, DAS, Ashok Kumar. Secure signature-based authenticated key establishment scheme for future IoT applications. IEEE Access, 2017, vol. 5, p. 3028-3043.

[5] HSIEH, Wen-Bin, and Jenq-Shiou Leu. "A robust user authentication scheme using dynamic identity in wireless sensor networks." Wireless personal communications 77.2 (2014): 979-989.

[6] DAS, Manik Lal. Two-factor user authentication in wireless sensor networks. IEEE transactions on wireless communications, 2009, vol. 8, no 3, p. 1086-1090.

[7] VAIDYA, Binod, MAKRAKIS, Dimitrios, MOUFTAH, Hussein T. Improved two-factor user authentication in wireless sensor networks. In: 2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications. IEEE, 2010. p. 600-606.

[8] F Wu, L Xu, S Kumari, and X Li "A privacy-preserving and provable user authentication scheme for wireless sensor networks based on internet of things security." Journal of Ambient Intelligence and Humanized Computing 8.1 (2017): 101-116.

[9] LI, Xiong, NIU, Jianwei, KUMARI, Saru. A three-factor anonymous authentication scheme for wireless sensor networks in internet of things environments. Journal of Network and Computer Applications, 2018, vol. 103, p. 194-204.

[10] JIANG, Qi, ZEADALLY, Sherali, MA, Jianfeng. Lightweight three-factor authentication and key agreement protocol for internet-integrated wireless sensor networks. IEEE Access, 2017, vol. 5, p. 3376-3392.

[11] Yeh, H. L., Chen, T. H., Liu, P. C., Kim, T. H., & Wei, H. W. (2011). A secured authentication protocol for wireless sensor networks using elliptic curves cryptography. Sensors, 11(5), 4767-4779.

[12] FENG, Qi, HE, Debiao, ZEADALLY, Sherali. Ideal lattice-based anonymous authentication pro ocol for mobile devices. IEEE Systems Journal, 2018, vol. 13, no 3, p. 2775-2785.

[13] FARASH, Mohammad Sabzinejad, TURKANOVIĆ, Muhamed, KUMARI, Saru, An efficient user authentication and key agreement scheme for heterogeneous wireless sensor network tailored for the Internet of Things environment. Ad Hoc Networks, 2016, vol. 36, p. 152-176.

[14] Amin, R., Islam, S. H., Biswas, G. P., Khan, M. K., Leng, L., & Kumar, N. (2016). Design of an anonymity-preserving three-factor authenticated key exchange protocol for wireless sensor networks. Computer Networks, 101, 42-62.

[15] Srinivas, Jangirala, Sourav Mukhopadhyay, and Dheerendra Mishra. "Secure and efficient user authentication scheme for multi-gateway wireless sensor networks." Ad Hoc Networks 54 (2017): 147-169.

[16] KHALIL, Nacer, ABID, Mohamed Riduan, BENHADDOU, Driss, Wireless sensors networks for Internet of Things. In: 2014 IEEE ninth international conference on intelligent sensors, sensor networks and information processing (ISSNIP). IEEE, 2014. p. 1-6.

[17] FARASH, Mohammad Sabzinejad, TURKANOVIĆ, Muhamed, KUMARI, Saru, An efficient user authentication and key agreement scheme for heterogeneous wireless sensor network tailored for the Internet of Things environment. Ad Hoc Networks, 2016, vol. 36, p. 152-176.

[18] Challa, Sravani, "Secure signature-based authenticated key establishment scheme for future IoT applications." IEEE Access 5 (2017): 3028-3043.

[19] Yeh, H. L., Chen, T. H., Liu, P. C., Kim, T. H., & Wei, H. W. (2011). A secured authentication protocol for wireless sensor networks using elliptic curves cryptography. Sensors, 11(5), 4767-4779.

[20] Shi, Wenbo, and Peng Gong. "A new user authentication protocol for wireless sensor networks using elliptic curves cryptography." International Journal of Distributed Sensor Networks 9.4 (2013):730831.

[21] TURKANOVIĆ, Muhamed, BRUMEN, Boštjan, HÖLBL, Marko. A novel user authentication and key agreement scheme for heterogeneous ad hoc wireless sensor networks, based on the Internet of Things notion. Ad Hoc Networks, 2014, vol. 20, p. 96-112.

[22] AMIN, Ruhul et BISWAS, G. P. A secure light weight scheme for user authentication and key agreement in multi-gateway based wireless sensor networks. Ad Hoc Networks, 2016, vol. 36, p. 58-80.

[23] Pauls, Friedrich, Robert Wittig, and Gerhard Fettweis. "A Latency-Optimized Hash-Based Digital Signature Accelerator for the Tactile Internet." International Conference on Embedded Computer Systems. Springer, Cham, 2019.

[24] Nakamoto, Satoshi, and A. Bitcoin. "A peer-to-peer electronic cash system." Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf (2008).

[25] STOJMENOVIC, Ivan et WEN, Sheng. The fog computing paradigm: Scenarios and security issues. In : 2014 federated conference on computer science and information systems. IEEE, 2014. p. 1-8.

[26] "Solidity", Accessed: Jan. 1, 2020. [Online]. Available: https://solidity.readthedocs.io/en/develop/.

[27] Ganache. Accessed: Jan. 1, 2019. [Online]. Available: https://truffleframework.com/docs/ganache/overview.

[28] Go Ethereum. Accessed: Mai. 1, 2019. [Online]. Available: https://geth.ethereum.org/.

PHY-DTR: An Efficient PHY based Digital Transceiver for Body Coupled Communication using IEEE 802.3 on FPGA Platform

Sujaya B.L[1]
Assistant Professor, Department of Electronics and Communication Engineering, BNMIT, Bengaluru, India

S.B. Bhanu Prashanth[2]
Professor, Department of Medical Electronics BMSCE, Bangalore, India

***Abstract*—Body coupled communication (BCC) is an efficient networking approach to body area network (BAN) based on Human-centric communication. The BCC provides interference only between humans in very close proximity. In this work, an efficient Physical layer (PHY) based digital transceiver) is designed for BCC. The digital transceiver Module mainly contains a Digital transmitter (TX) with Manchester encoder, clock synchronization unit, and Digital receiver (RX) with Manchester decoder. The TX and RX modules are designed using a finite state machine as per the IEEE 802.3 Standards. The complete work is also varied for BAN applications by connecting two Application layer transceivers and two Physical layer-based digital transceivers. The architecture is simulated in a Model-sim simulator. The complete Module is synthesized using different FPGA families, and the hardware design constraints are contrasted. The digital transceiver works at 231.28 MHz operating frequency, consumes 0.113W power, and provides a 7.7 Mbps data rate and 4.67 Kbps/Slice efficiency on Artix-7 FPGA. The proposed transceiver is also compared with existing digital transceivers with hardware constraints improvements.**

***Keywords*—Body coupled communication; physical layer; digital; FPGA; radiofrequency; human body**

I. INTRODUCTION

The body area network technology with wireless connectivity is one of the promising technologies in the Health care domain because of its flexibility and portability. The BAN visualizes and controls the operating sensors, which can measure the critical physiological and physical parameters. Radiofrequency (RF) based wireless technology is deployed in BAN systems, which lag battery power, security, and electromagnetic issues [1]. Intra body-communication (IBC) is an alternative non-RF technology, which uses the human body as a transmission medium for signals (electrical) and overcomes most of the RF-technology-related issues [1-2]. The IBC gives lower power and better data rates; this results in an alternative to short-range communications. The IBC has different types, each having unique properties. The well-suited wireless communications types are Ultrasound (US), galvanic coupling, resonant coupling, and capacitive coupling [3-4]. There are many electrical coupling techniques available for data transmission through the human body. In general, electromagnetic wave and electrostatic coupling are the two commonly used transmission methods in IBC. The IBC is used in many applications like touching voice system, blind person assistance system, speech assistance for dumb persons [5]. The wireless BAN comprises mainly three PHY schemes: Narrowband (NB), ultra-wideband (UWB), and Human body communication (HBC) as per IEEE 802.15.6 standardization [6]. The NB and UWB schemes are related to RF technologies, whereas HBC is related to Non-RF technologies. The HBC is also considered body channel communication and body coupled communications in the current work [7]. The human body is used as a signal transmission medium in BCC, connected with electronic devices nearby, to provide human-centric communication. The BCC transceiver [8-9] is designed to achieve better power efficiency and data rates by avoiding the fading effects. The Overview of the BCC transceiver architecture is represented in Fig. 1. It mainly contains BCC TX and RX modules. The BCC-TX mainly has three parts: Application layer, PHY-based digital TX, and Analog front-end (AFE) part. Similarly, BCC-RX has an AFE part, followed by a PHY-based digital RX and application layer.

The Digital Transmitter receives data from the application layer (AL) and converts it into a packet sent using unified communication over the human body. These packets are processed further in the AFE part of the transmitter. The Digital receiver collects the data serially from the AFE receiver part and recovery the data with proper operation in DR. It sends it back to the application layer. The complete PHY-based D-TR is working in full-duplex communication mode for BCC. The application layer collects data from the user on the transmitter side and displays its output. The Ethernet protocol is used in PHY as per IEEE 802.3 standards for transmitting and receiving BCC transceiver data. The Manchester encoding and decoding modulation techniques are incorporated in the transceiver, which is used significantly in the human body with capacity coupling. The transmission line or capacitive approaches are used as a coupling mechanism, interconnecting BCC TX and BCC RX of the Human body. There is always a challenge to design communication protocols for BCC by analyzing each layer's -fundamental features like the application layer, Media Access Control (MAC) layer, and physical layer. The MAC layer supports the IEEE 802.3 ethernet Protocol, which provides a high data rate, and more influenced for real-time application cases for the BBC system.

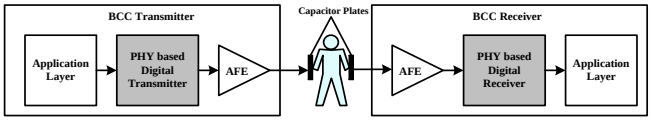


Fig. 1. BCC Transceiver Architecture Overview.

The PHY-based digital transceiver architecture is designed for BCC and Prototyped on the FPGA platform in this work. The proposed transceiver provides effective data rates and that can be used for Human Body. The rest of the paper is organized as Section II explains the existing BCC works and its transceiver modules using different design parameters. The proposed PHY-based digital Transceiver architecture is explained in detail in Section III. Section IV discusses the different design constraints of PHY-based digital transceivers on different FPGA families. Finally, Section V concludes the overall work of the PHY-based digital transceiver and suggests the future scope.

II. Related Work

In this section, the review of the existing body coupled communication architectures is done. Arenas et al. [10] present the BCC module for streaming music using Universal software Radio peripheral (USRP) hardware module and GNU radio. The signal generator is used to measure the body channel frequency and its response. The obtained 128 kbps data rate is verified using a Vector network analyzer (VNA). Achieving a higher data rate is noted to be difficult in this work. Yoo et al. [11] discuss the energy-efficient BCC for Body area network applications. The work analyses the space, time concerning channel gain, environmental variation, power consumption details, and benefits. The work uses a pseudo OFDM transceiver module, which consumes more chip area and power in real-time implementations. Kado et al. [12] present the embedded transceiver architecture for the human body. The work is based on near field coupling architecture using human Area networks (HAN). The work discusses the better packet error rates (PER) by concerning the received signal power for uplink and downlink transmission. The embedded transceiver achieves minimal data rate and consumes more power for received packets. Matias et al. [13] discuss the Capacitive BCC via the ECG acquisition system. This work elaborates on capacitive coupling with the BCC transceiver system connected to the ECG monitoring system. It analyses the received ECG signal with a 100kbps data rate on the mobile device. The designed BCC system is suitable only for ECG monitoring applications and not applicable for HBC. Takeuchi et al. [14] present the wearable Near-field coupling (NFC) Transceiver architecture using handshaking communication. The battery-powered TR with a digital synthesizer is an equivalent circuit for handshaking communication between two humans. The results of signal propagation loss and received voltage concerning two human bodies are analyzed. The work is suitable only if humans were 600-800 mm apart and is reported to be not suitable for higher ranges.

Saadeh et al. [15-16] implement the BCC TR for Binaural hearing aids for BAN. This work uses Pseudo OFDM TR with Frequency shifting keying schemes for BCC TR to improve TR's Bit error rate at a 1 Mbps data rate. The traditional Pseudo OFDM TR consumes more chip area and power and not suitable for real-time HBC. Zhao et al. [17] present Human Body communication TR compatible with IEEE 802.15.6 and achieves less BER and 5.25 Mc/s chip rate. The mask-shaped transmitter and digital controlled calibration-based receiver are designed for HBC TR. The designed work is intricate and consumes more area on the 65nm- CMOS chip process. Chung et al. [18] implement the BCC TR using Walsh codes on the FPGA platform for HBC. The jitter tolerance and code rate are improved using Walsh codes. The BCC TR achieves the 10-8 BER at a 6.25 Mcps data rate. The implemented BCC TR uses only Walsh code-based data transmission without using MAC features.

Park et al. [19] discuss the Magnetic HBC TR by enabling 5Mbps data at 40MHz carrier frequency. This work discusses the design challenges of different HB TR modulation and demodulation schemes. The results are obtained on the ASIC Platform and analyze the data rate and power consumption. The data rate can be further improved by using a suitable modulation scheme. Muzaffar et al. [20] present the BCC TR, which provides low-power, self-synchronization, and low complexity while implementing on the human body. The TR is verified on the oscilloscope for received and transmitted signals through the body. The work also analyses the energy and power consumption of BCC TR with an average data rate of 21Mbps at 125MHz. Krhac et al. [21] present the HBC channel analysis on the simulation platform, which provides a forward transmission coefficient for various capacitive return paths and electrode distances. The work is analyzed in the software environment and not compactable to real-time HBC. Jeon et al. [22] discuss the BCC TR for Bionic arms using a galvanic coupling. The BCC TR is designed using CMOS 0.18 µm, which provides better energy efficiencies of 4.75 pJ/b and 26.8pJ/b for TX and RX, respectively, and improves the 10-8 BER using a galvanic coupling at 100Mbps data rate. Yoo et al. [23] present the BAN TR using BCC, which provides a better energy-efficient TR communication system. The conventional pseudo –OFDM-based BCC TR is designed for BAN, which offers a 1Mbps data rate.

Wei et al. [24] present the intra-body communication (IBC) TR with galvanic coupling (GC) using differential phase shifting keying (DPSK) schemes. This work achieves a 1Mbps data rate with 0.6mA of coupling amplitude. The data acquisition module is designed in a Lab-view environment, and TR is designed on FPGA, which consumes more chip area. Chen et al. [25] discuss the GC-based IBC TR using Direct sequence spread spectrum (DS-SS) Technology. The DSSS-DPSK based TR achieves better BER and SNR than DPSK based TR. The designed TR is complex and works at a 50kbps data rate, and will not effective in WBAN. Botero et al. [26] review the HBC channel characteristics' issues using different coupling techniques with different measurement approaches. Slot et al. [27] present the heartbeat-based MAC architecture for BCC applications. The TDMA based MAC protocol is introduced for packet transmission and reception in capacitive BCC architecture for heartbeat sensing. Ormanis et al. [28] discuss human body frequency response as a case study in BCC for e-Health. The human body frequency

response range is up to 30MHz using BCC. Li et al. [29] present the Differential AFE receiver for Galvanic-coupled HBC. The Stability of AFE in the frequency range is up to 1MHz, and AFE reduces the interference margin to a great extent. Vizziello et al. [30] present PHY implementation for IBC links using a galvanic coupling. The Galvanic coupling method analyses the selection of modulation schemes, frequency parameters, and recovery of baseband signals.

It has been noticed from the above existing work that most of the BCC transceivers are entirely designed on software and very few on FPGA and ASIC Platform. Significantly less work on the PHY-based digital transceiver with MAC features support for BCC system. There remains a scope to develop schemes to achieve higher data rates with novel transceiver architecture designs. In this proposed work, an efficient PHY-based digital DTR is designed to overcome data rate limitations.

III. PHY BASED DIGITAL TRANSCEIVER

The physical layer-based D-TR is designed for BCC and is detailed in this section. The IEEE 802.3 Ethernet protocol standard is adopted for transmitting and receiving the packets in PHY-based D-TR. The IEEE 802.3 Ethernet protocol supports the 10Mbps data rate with the baseband signaling method, and the co-axial medium is selected in PHY. The PHY-based D-TR module contains a digital transmitter, digital receiver, Manchester encoder and decoder, and clock synchronization module (CSM), as represented in Fig. 2. The PHY-based digital TR sub-modules are explained in the sub-sections next.

A. Digital Transmitter (D-TX)

The AL provides the data information to the D-TX and proceeds with Manchester encoded data serially in the transmitter's AFE part. The D-TX mainly contains two memory modules-one for preamble generation and another for data storage. The D-TX has a Packet framing unit (PFU), shift register, Frame check sequence (FCS) unit, Transmitter-FSM for controlling the D-TX, and Manchester encoder. The D-TX clock frequency is set at 100MHZ. The D-TX receives the data from the application layer, temporarily stores the data in memory, and assembles the packet form data. It is represented in Fig. 3. The preamble generation information is stored in memory 1 (MEM1). The MEM1 holds eight 8-bit preamble values and is used for clock synchronization with PHY. The Packet framing unit processes the application layer data as per the IEEE 802.3 standard.

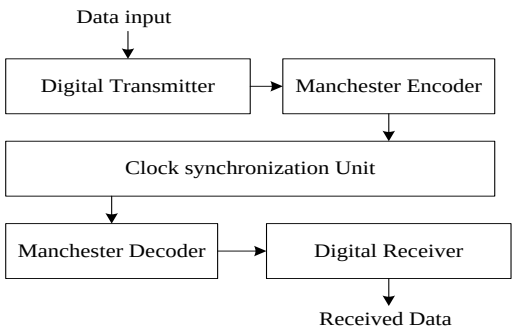


Fig. 2. PHY based Digital Transceiver (D-TR) Architecture.

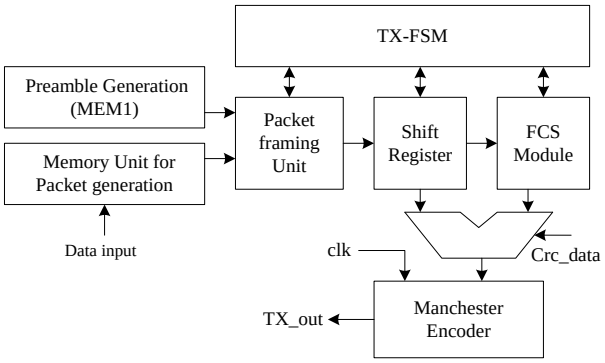


Fig. 3. PHY based Digital Transmitter Architecture.

The IEEE 802.3 Packet format for Ethernet protocol is represented in Fig. 4. The Packet format starts with a preamble, followed by the frame (SOF) Delimiter, Length field, Source and destination ID, payload data, and Frame check sequence.

The 7-bytes preamble is used to synchronize the clock with the physical layer and intimates incoming frames to the receiver station. The SOF delimited initiates the first-byte frame data. The 2-6 bytes of Source and destination ID contains address information of transmitting and receiving station, respectively. The 2-byte length field provides payload (data) information. The 46-1500 bytes of the payload can be accessed in the Ethernet frame. The 2-4 bytes of frame check sequence provides error detection features using cyclic redundancy check (CRC). In the D-TX module, 7-Bytes of Preamble, 1-byte of SOF delimiter, 1-byte of source and destination ID, 2-bytes of length, 64-bytes of payload data, and 2-bytes of FCS data is considered.

The transmitter-FSM is mainly used to control switching activity between the different fields and provides proper communication. The detailed TX-FSM is represented in Fig. 5.

7-bytes	1-byte	2-6 bytes	2-6 bytes	2-bytes	46-1500 bytes	2-4 bytes
Preamble	SOF	Destination ID	Source ID	Length	Payload	FCS

Fig. 4. Standard IEEE 802.3 Packet format (Ethernet).

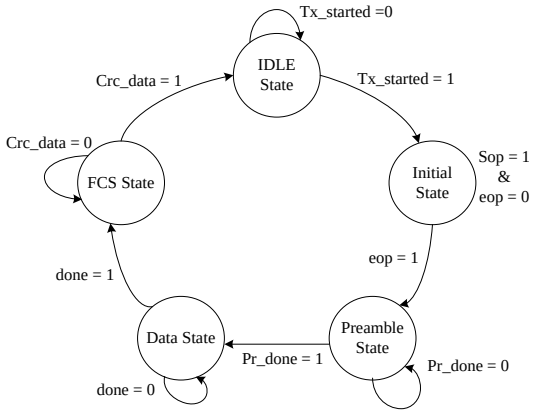


Fig. 5. Transmitter-FSM Diagram.

The five states are used in FSM, IDLE state, Initial state, Preamble state, data state, and FCS state. In the IDLE state, the transmission process (TX started) is started when RX is not busy. The initial state provides first Packet data by providing two handshaking signals (start of packet (sop) =1 and end of packet (eop) =0) from the PHY medium. When eop is activated, the preamble state sends the preamble data by activating Pr_done=1 to the data state. In the data state, the input data receiving from the memory unit and activates the done signal. In the FCS state, the CRC is used to check the packets' status, like packet data is valid or not; the packet is short or long; any error occurs in the packet or not. The CRC has also detected errors and validate the packet. The shift register is used to generate the serial data by shifting 1-bit left using a framed packet. The FCS module receives the serial data and checks the error status using CRC. If the CRC data is activated, the FCS data else shift register's serial data is used in the Manchester encoder. The CRC is modeled using a linear feedback shift register (LFSR). The 16-bit LFSR polynomial is used for error detection for the given packet. The 16-bit LFSR polynomial for CRC module is expressed as: 1 + x4 + x11+ x15. The Manchester encoder provides better baseband modulation to the D-TX module.

The clock signal is XOR with serial data for the generation Manchester encoded data (Tx_data) as a D-TX output. The Encoded data is in the form of '0' or '1'. The transition occurs either from '0' to '1' and from '1' to '0' in every clock cycle. In further, these encoded data are used in the clock synchronization unit to recover the clock.

B. Clock Synchronization Unit

The clock synchronization unit receives the encoded data serially in an oversampling manner. The clock synchronization unit uses the clock signal 8 times faster than the D-TX clock signal. This clock signal provides the transition edges by using the receiving the data signal. If the Manchester encoded data is one, the clock detects it as a positive edge; otherwise, it is a negative edge. The combination of positive and negative edges detects the recovered clock when similar data bits match. If the received data is not matched, then the central transition signal is activated, with the clock signal's absence in the received data. The recovered clock is oversampled with received data for the formation of the decoded output. The Manchester decoder uses X-NOR operation to decodes the received data with the recovered clock. The decoded data is used as a serial input to the D-RX module.

C. Digital Receiver

The PHY-based Digital Receiver Architecture and The RX-FSM are represented in Fig. 6 and Fig. 7, respectively. The D-RX mainly contains a Delay unit, Preamble checking unit, SOF finding unit, length activation unit, packet recovery unit, matching unit, Memory unit, and RX-FSM. The D-RX receives the Manchester decoded data and input it into the delay unit. The delay unit contains many data flip-flops (D-FF), which synchronize the decoded data for proper packet recovery. The RX-FSM is used to control all the submodules with proper interconnection signals.

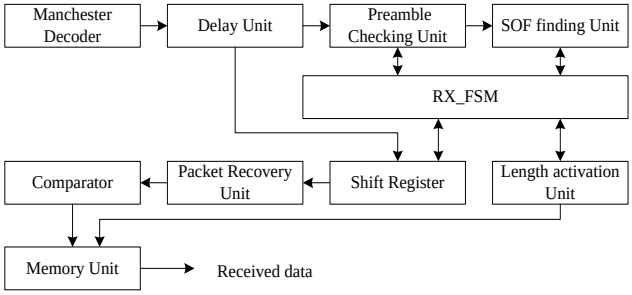


Fig. 6. PHY based Digital Receiver Architecture.

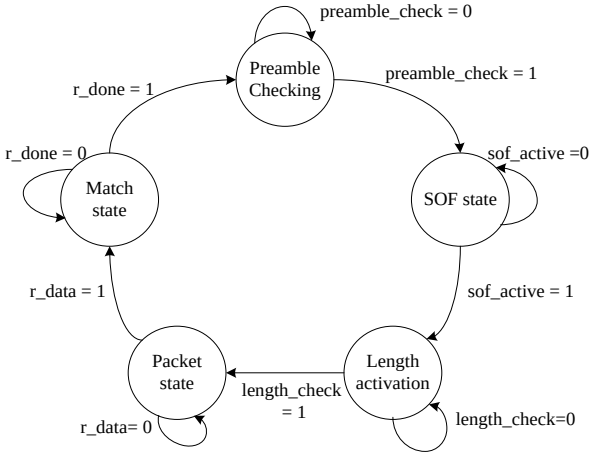


Fig. 7. Receiver-FSM Diagram.

The RX-FSM mainly contains five states, namely: Preamble checking state, SOF state, length activation state, packet recovery state, match state. The Preamble checking state checks the first 16-bits of the data; if it matches, then move to the SOF delimiter state. In SOF State, the two successive '1' appear, then the state transitions to the Length activation state. Decode the first two bytes of the data using a shift register and delay unit to check the packet's length and store these two bytes of the packet length in the Memory unit. The packet state recovers the data packet after the recovery of the complete data from the delay unit. The match state checks the source ID with the next consecutive of the recovered packet, if it matches, the recovery operation is completed, and signal r_done is activated. The D-RX Module Stop receiving data from the clock synchronization unit. The recovered data stored in memory to validate the received and transmitted data packets are not. If the recovered packets are not matched, then the same process repeats until valid packets are recovered.

The designed PHY-based digital transceiver is verified for Body area network applications, represented in Fig. 8. Two Application layer transceivers (AL-TR1 and AL-TR2) and Two PHY-based digital transceivers are used for verification. The AL-TR1 received the data (Tx) from the user, passed to the PHY-TR1, and received the output data (Rx) from PHY-TR1. Similarly, The AL-TR2 received the data (Tx) from the user, passed to the PHY-TR2, and received the output data (Rx) from PHY-TR2. The standard clock signal is used for two PHY-TR, which is received from the AL-TR. The clock

signal frequency may vary upon application usage. For example, when the PHY-TR1 wants to send the data to PHY-TR2, it first checks the availability of PHY-TR2. If it is ready, the data is sent to the PHY-TR2, which decodes the data packet until the destination ID matches and processes further to the AL-TR-2.

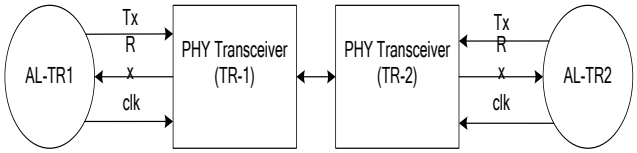


Fig. 8. PHY-Transceiver Verification overview for BAN Application.

IV. Results and Discussion

The results of the PHY-based digital transceiver are discussed in this section. The Complete architecture is designed using VHDL on the Xilinx ISE environment and simulated using a Modelsim simulator. The PHY-based digital transceiver module is implemented and prototyped using Artix-7 FPGA (XC7A100T-3CSG324). The PHY-based digital transceiver submodules design constraints are tabulated in Table I. The Digital TX utilizes 750 slices, works at 297.38 MHz, and consumes 0.088W total power.

Similarly, the clock synchronization (Clock_Sync) unit utilizes 45 slices, works at 352.78 MHz, and consumes 0.104W total power. The clock synchronization module uses an 800MHz clock frequency, which is 8-times faster than the Digital TX clock frequency. The Digital RX utilizes 696 slices, works at 219.13 MHz, and consumes 0.090W total power.

The Performance parameters for PHY-based Digital transceivers using different FPGA families like Spartan-6, Artix-7, and Viretx-7 are tabulated in Table II. The Spartan-6 FPGA uses 45-Nm CMOS Technology, whereas Artix-7 and Virtex-7 use 28nm CMOS Technology. The PHY-based D-TR utilizes 1672 slices, 2586 LUT's, works at 145.21 MHz, and consumes 0.092W total power on Spartan-6 FPGA. Similarly, The PHY-based D-TR utilizes 1646 slices, 2517 LUT's, works at 231.28 MHz, and consumes 0.113W total power on Artix-7 FPGA. The PHY-based D-TR utilizes 1646 slices, 2518 LUT's, works at 310.01 MHz, and consumes 0.22W total power on Virtex-7 FPGA.

The PHY-based Digital Transceiver latency is calculated based on the simulation results obtained using the Modelsim simulator. The PHY-based Digital Transceiver takes 240 clock cycles (CC) to receive the first packet data. The period of one clock cycle is defined as 16 ns. The PHY-based Digital Transceiver throughput is calculated using latency (CC), the number of input data passed, and the design's maximum operating frequency. So, throughput (Mbps) = (Input data * Max. frequency)/latency. Only 7 input packets are considered for simulation purposes. The throughput (data rate) of 4.84Mbps, 7.7Mbps, and 10.33Mbps obtained on Spartan-6, Artix-7, and Virtex-7 FPGA families PHY-based Digital Transceiver. The *Throughput of 3.33 Mbps is obtained for Maximum operating frequency 100MHz, a suitable BCC Transceiver [9] Module. The Hardware efficiency is calculated based on Throughput per slice (Kbps/Slice). The Hardware efficiency of 2.89 Kbps/Slice, 4.67 Kbps/Slice, and 6.27 Kbps/Slice was obtained for Spartan-6, Artix-7 Virtex-7 FPGA families, respectively.

The Chip Area utilized for PHY-based Digital Transceiver on different FPGA families like Spartan-6, Artix-7, and Viretx-7 are represented in Fig. 9(a), 9(b), and 9(c), respectively. The Chip Area Utilization on different FPGAs is obtained after the place and route operation using the Xilinx FPGA editor Tool.

TABLE I. Resources utilized for PHY based Digital Transceiver Sub-Modules on Artix-7 FPGA

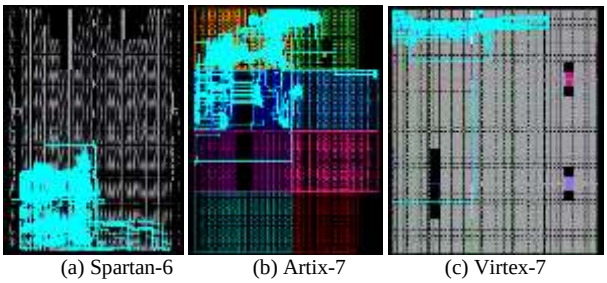
Resource Used	Digital TX	Clcok_Sync Unit	Digital RX
Slice Registers	750	45	696
Slice LUTs	760	99	1129
LUT-FF pairs	230	43	662
Minimum Period (ns)	3.363	2.835	4.563
Max.Frequency (MHz)	297.38	352.787	219.13
Total Power (W)	0.088	0.104*	0.09

*800 MHz clock Frequency (which is 8 times faster than TX clock frequency)

TABLE II. Performance Parameters for PHY based Digital Transceiver using different FPGA's

Resource Used	PHY based Digital Transceiver on Different FPGA Families		
FPGA Family	Spartan-6	Artix-7	Virtex-7
FPGA Device	XC6SLX45T-3CSG324	XC7A100T-3CSG324	XC7V330T-3FFG1157
CMOS Technology	45nm	28nm	28nm
Slices	1672	1646	1646
LUTs	2586	2517	2518
Max. Frequency (MHz)	145.214	231.28	310.017
Total Power (W)	0.092	0.113	0.22
Latency (Clock cycles)	240	240	240
Throughput (Mbps)	4.84	7.7	10.33
*Throughput (Mbps)	3.33	3.33	3.33
Efficiency (Kbps/Slice)	2.89	4.67	6.27
*Efficiency (Kbps/Slice)	1.91	2.02	2.02

*100 MHz Operating Frequency suitable for BCC [9]



(a) Spartan-6 (b) Artix-7 (c) Virtex-7

Fig. 9. Chip Area utilized for PHY based Digital Transceiver on different FPGA.

The comparison results of different PHY based Digital Transceiver's on hardware platform is tabulated in Table III. For the comparison, different hardware constraints like the selection of FPGA device, Slice, LUT's Power (mW), and throughput (Mbps) are considered. The baseband transceiver [31] is implemented on spartan -6 FPGA for BAN communication. The work utilizes 2843 slice registers, 3915 LUT's, consumes 192mW power, and works at 0.187Mbps. The Proposed digital transceiver is compared with the existing transceiver [31], with an improvement in overhead for 41.18% in slices, 33.94% in LUT's and 52% in power utilization. The physical transceiver [32] is designed on Virtex-6 FPGA for WBAN applications, which utilizes 2668 slice registers, 3161 LUT's, consumes 117mW power, and works at 0.121 Mbps. The Proposed digital transceiver is compared with the existing PHY transceiver [32], with an improvement in overhead for 37.33% in slices, 18.19% in LUT's and 21.3% in power utilization. The proposed PHY-based digital transceiver provides better throughput than the other two compared transceivers.

The proposed digital transceiver is designed, and it is optimized with the help of FSM's. The proper data packets transmission and reception are achieved using TX and RX-FSM's with a clock synchronization mechanism. The FSM is activated only when necessary or in the initial state while transmission/ reception of data packets.

TABLE III. COMPARISON RESULTS OF DIFFERENT PHY BASED DIGITAL TRANSCEIVER'S

Resources	**Ref [31]**	**Ref [32]**	**Proposed**
FPGA	Spartan-6	Virtex-6	Spartan -6
Slices	2843	2668	1672
LUT's	3915	3161	2586
Power (mW)	192	117	92
Throughput (Mbps)	0.187	0.121	4.84

V. CONCLUSION

In this paper, an efficient PHY-based D-TR is designed for BCC. The PHY-based D-TR receives or transit the packets to and from the application layer and supports full-duplex combination using 802.3 IEEE communication protocol standards. The PHY-based D-TR incorporates Manchester encoding-decoding mechanism to recover the complete data packet with proper clock synchronization. The architecture utilizes 1% slices, 3% LUT's, works at 231.28 MHz maximum frequency, and consumes a total power of 0.113W on Artix-7 FPGA. The architecture works at 7.7 Mbps throughput with an efficiency of 4.67 kbps/slice on Artix-7 FPGA. The PHY-based Digital Transceiver synthesized on different FPGA families like Spartan-6, Artix-7, and Virtex-7 and obtained different design constraint results. The PHY-based D-TR obtains a throughput of 3.33Mbps at 100 MHz, suitable for the BCC [9] system. In the future, the design of a complete BCC transceiver using the proposed PHY-based digital transceiver design of Human body communication transceiver can be modeled as per 802.15.6 IEEE standard for BAN applications.

REFERENCES

[1] Seyedi, M., Kibret, B., Lai, D. T., & Faulkner, M. (2013). A survey on intrabody communications for body area network applications. IEEE Transactions on Biomedical Engineering, 60(8), 2067-2079.

[2] Zhao, J. F., Chen, X. M., Liang, B. D., & Chen, Q. X. (2017). A review on human body communication: signal propagation model, communication performance, and experimental Issues. Wireless Communications and Mobile Computing, 2017.

[3] Tomlinson, W. J., Banou, S., Yu, C., Stojanovic, M., & Chowdhury, K. R. (2018). A comprehensive survey of galvanic coupling and alternative intra-body communication technologies. IEEE Communications Surveys & Tutorials, 21(2), 1145-1164.

[4] Corroy, S., Argyriou, A., Bhatti, Z. W., & Baldus, H. (2010). A body-coupled communication and radio frequency dual technology cooperation protocol for body area networks. In ICC'10 Workshop on Medical Applications Networking, Date: 2010/05/27-2010/05/27, Location: Cape Town, SA. IEEE.

[5] Shimamoto, S., Alsehab, A. M., Kobayashi, N., Dovchinbazar, D., & Ruiz, J. A. (2007, December). Future applications of body area communications. In 2007 6th International Conference on Information, Communications & Signal Processing (pp. 1-5). IEEE.

[6] Astrin, A. (2012). IEEE standard for local and metropolitan area networks part 15.6: Wireless body area networks. IE EE Std 802.15. 6.

[7] Song, S. J., Cho, N., Kim, S., Yoo, J., Choi, S., & Yoo, H. J. (2007, February). A 0.9 V 2.6 mW body-coupled scalable PHY transceiver for body sensor applications. In 2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers (pp. 366-609). IEEE.

[8] Fazzi, A., Ouzounov, S., & van den Homberg, J. (2009, February). A 2.75 mW wideband correlation-based transceiver for body-coupled communication. In 2009 IEEE International Solid-State Circuits Conference-Digest of Technical Papers (pp. 204-205). IEEE.

[9] Baldus, H., Corroy, S., Fazzi, A., Klabunde, K., & Schenk, T. (2009). Human-centric connectivity enabled by body-coupled communications. IEEE Communications Magazine, 47(6), 172-178.

[10] Arenas, G. M., & Gordillo, A. C. (2016, October). Design and implementation of a body-coupled communication system for streaming music. In 2016 IEEE ANDESCON (pp. 1-4). IEEE.

[11] Yoo, J. (2017). Body coupled communication: Towards energy-efficient body area network applications. In 2017 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT) (pp. 244-246). IEEE.

[12] Kado, Y., Kobase, T., Yanagawa, T., Kusunoki, T., Takahashi, M., Nagai, R & Shinagawa, M. (2012, January). Human-area networking technology based on near-field coupling transceiver. In 2012 IEEE Radio and Wireless Symposium (pp. 119-122). IEEE.

[13] Matias, R., Cunha, B., Mota, A., & Martins, R. (2014, June). ECG monitoring via Capacitive Body Coupled Communications. In 2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6). IEEE.

[14] Takeuchi, R., Hasegawa, S., Kado, Y., Ayuzawa, D., Shinagawa, M., Ohashi, K., & Saito, D. (2017, March). Implementation methodology of handshaking communication using wearable near-field coupling transceivers. In 2017 11th European Conference on Antennas and Propagation (EUCAP) (pp. 1871-1875). IEEE.

[15] Saadeh, W., Altaf, M. A. B., Alsuradi, H., & Yoo, J. (2017). A pseudo OFDM with miniaturized FSK demodulation body-coupled communication transceiver for binaural hearing aids in 65 nm CMOS. IEEE Journal of Solid-State Circuits, 52(3), 757-768.

[16] Saadeh, W., Altaf, M. A. B., Alsuradi, H., & Yoo, J. (2017). A 1.1-mW ground effect-resilient body-coupled communication transceiver with pseudo OFDM for head and body area network. IEEE Journal of Solid-State Circuits, 52(10), 2690-2702.

[17] Zhao, B., Lian, Y., Niknejad, A. M., & Heng, C. H. (2018). A low-power compact IEEE 802.15. 6 compatible human body communication transceiver with digital sigma-delta IIR mask shaping. IEEE Journal of Solid-State Circuits, 54(2), 346-357.

[18] Chung, C. C., Chang, R. H., & Li, M. H. (2018, May). An FPGA-Based Transceiver for Human Body Channel Communication Using Walsh

Codes. In 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) (pp. 1-2). IEEE.

[19] Park, J., & Mercier, P. P. (2019). A Sub-10-PJ/bit 5-Mb/s Magnetic Human Body Communication Transceiver. IEEE Journal of Solid-State Circuits, 54(11), 3031-3042.

[20] Muzaffar, S., & Elfadel, I. M. (2019, July). A Self-Synchronizing, Low-Power, Low-Complexity Transceiver for Body-Coupled Communication. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4036-4039). IEEE.

[21] Krhac, K., Sayrafian, K., Noetscher, G., & Simunic, D. (2019, July). A Simulation Platform to Study the Human Body Communication Channel. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4040-4043). IEEE.

[22] Jeon, Y., Jung, C., Cheon, S. I., Cho, H., Suh, J. H., Jeon, H., & Je, M. (2019, June). A 100Mb/s Galvanically-Coupled Body-Channel-Communication Transceiver with 4.75 pJ/b TX and 26.8 pJ/b RX for Bionic Arms. In 2019 Symposium on VLSI Circuits (pp. C292-C293). IEEE.

[23] Yoo, J. (2019). Energy-Efficient Body Area Network Transceiver Using Body-Coupled Communication. In The IoT Physical Layer (pp. 127-139). Springer, Cham.

[24] Wei, Z. L., Chen, W. K., Yang, M. J., Gao, Y. M., Vasić, Ž. L., & Cifrek, M. (2020, May). Design and Implementation of Galvanic Coupling Intra-Body Communication Transceivers using Differential Phase Shift Keying. In 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) (pp. 1-6). IEEE.

[25] Chen, W. K., Wei, Z. L., Gao, Y. M., Vasić, Ž. L., Cifrek, M., Vai, M. I & Pun, S. H. (2020). Design of Galvanic Coupling Intra-Body Communication Transceiver Using Direct Sequence Spread Spectrum Technology. IEEE Access, 8, (pp 84123-84133).IEEE.

[26] Álvarez-Botero, G. A., Hernández-Gómez, Y. K., Telléz, C. E., & Coronel, J. F. (2019). Human body communication: Channel characterization issues. IEEE Instrumentation & Measurement Magazine, 22(5), (pp. 48-53).

[27] Solt, F., Benarrouch, R., Tochou, G., Facklam, O., Frappé, A., Cathelin, A & Rabaey, J. M. (2020). Energy Efficient Heartbeat-Based MAC Protocol for WBAN Employing Body Coupled Communication. IEEE Access, 8, 182966-182983.

[28] J. Ormanis and A. Elsts, (2020). Towards Body Coupled Communication for eHealth: Experimental Study of Human Body Frequency Response. In 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, (pp. 1-7). IEEE.

[29] Li, D., Wu, J., Gao, Y., Du, M., Vasić, Ž. L., & Cifrek, M. (2020, May). A Differential Analog Receiver Front-End for Galvanic-Coupled Human Body Communication. In 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) (pp. 1-5). IEEE.

[30] Vizziello, A., Savazzi, P., Magenes, G., & Gamba, P. (2020). PHY Design and Implementation of a Galvanic Coupling Testbed for Intra-Body Communication Links. IEEE Access, 8, 184585-184597.

[31] Liang, Y., Zhou, Y., & Li, Y. (2014). The design and implementation of IEEE 802.15. 6 Baseband on FPGA. In The International Conference on Health Informatics (pp. 231-235). Springer, Cham.

[32] Mathew, P., Augustine, L., Kushwaha, D., Desalphine, V., & Selvakumar, A. D. (2015, January). Implementation of NB PHY transceiver of IEEE 802.15. 6 WBAN on FPGA. In 2015 International Conference on VLSI Systems, Architecture, Technology, and Applications (VLSI-SATA) (pp. 1-6). IEEE.

Towards an Ontological Learner's Modeling During and After the COVID-19 Pandemic

Amina OUATIQ[1], Kamal El Guemmat[2], Khalifa Mansouri[3], Mohammed Qbadou[4]
SSDIA Laboratory, ENSET Mohammedia
Hassan II University of Casablanca
Mohammedia, Morocco

***Abstract*—The health crisis and the unprecedented upheaval in the education systems which it caused are far from being over, consequently, the adaptation of the learning experience is most needed, and it should take into consideration the criteria of this specific crisis and its impact on the physical and mental health of the learners. In this article, we aimed to present an ontology-based learner model that will bring together the pedagogical and psychological characteristics, but also the health risks generated by the epidemic on the learners, following a Knowledge-Engineering Methodology. We developed an ontology that combines the IMS-LIP standard features and the learner characteristic. It is ready for different uses in different systems and situations during and after the COVID-19 pandemic, and it will give a global representation of the learner in order to allow him to get the best-adapted courses.**

Keywords*—Learner model; personalization; adaptive learning; ontology; COVID-19

I. Introduction

With the pandemic and the closure of schools and universities, in order to control the spread of the virus. Different distance learning approaches have been developed to ensure the continuity of education. The transition from face-to-face to distance education with little to no experience has generated an educational crisis in addition to the health and economic crisis, especially in countries with a low Human Development Index [1].

Even before the pandemic, teachers always tried to prepare and offer adapted courses to their learners, taking into consideration that learners are the center of the learning experience, and that every learner is special and learns distinctively according to his learning style, cognitive style, previous knowledge, as well as his behavior and motivation [2]. Whether in hybrid, or distance education adaptive systems are the best tools to provide learners with the best learning objects based not only on their preferences but also on their needs noting that learners themselves might not even be aware of their own needs, particularly in times of crisis such as this pandemic.

Adaptive systems utilize individual informations available on learners, in order to offer a specific and accurate learning experience to each learner. Therefore, the learner model should be the most radical part because it helps to understand the learner and give him an active role in his learning [3].

There are several ontological learner models that discuss various characteristics such as the learner's learning style, performance [4-7], motivation [8], knowledge [6, 9]. However, during the pandemic, and being in lockdown, other characteristics related to COVID-19 and to learner's mental health has emerged. Yet there is no model that addresses those characteristics or the impact of the crisis time on the mental state of learners.

The absence of learner models that take into account the learners' mental health and the impact of times of crisis on their well-being is the problem that this article tries to solve by developing an ontological learner model that represent relevant pedagogical, psychological, and physical characteristics, using the health crisis related to the COVID-19 pandemic and the resulting isolation impacts.

This article will present a learner model ontology, which will not only feature the characteristics of each learner such as knowledge, background, goals, interests, learning styles, and learning activities [10], but also other properties relevant to the time of the pandemic, and helpful to cope with the effects of confinement on the mental health of learners, as if the learner or a member of his family is affected by COVID-19, the presence of the symptoms of anxiety and/or depression, as well as if he lives in a cluster site. The proposed ontology will allow the reuse of data, it is evaluated and structurally validated [11], so all the requirements and constraints are met to make use of it.

The rest of this article is organized as follows. The following section will present the standard learner models and their attributes, and the effects of COVID-19 on the learner's mental health. It will then describe the methodology of designing the ontology in the third section. And in the fourth section, the learner model ontology will be presented. Later in the fifth section some examples of uses of the ontology will be suggested. And will conclude in the last section.

II. Related Work

A. Learner Model

The learner model is a representation of cognitive and non-cognitive characteristics of the learner. This model allows the representation and management of each learner with his own individual properties and characteristics [12, 13].

The learner model is different from the learner profile which only contains static data. While the model contains both static and dynamic data [12]. So, the profile can be the basis of the model, which is a more like an abstract representation, and the ontologies allow to define a common vocabulary on an abstract domain, and to specify the complex relations between the concepts of domain [14]. This is among the reasons they have become one of the most widely used learner modeling approaches [15], and due to its advantages, such as ease of reuse, availability of effective design tools, etc. [16].

The purpose of learner modelling is to give a complete and accurate description of all aspects of the user's behavior [17]. User models in adaptive hypermedia are generally compatible with learner standards. There are several standard learner models like IMS ACCLIP (Accessibility for Learning Information Package), IMS RDCEO (Reusable Definition of Competency or Educational Objective), and others but the most used are PAPI learner (Public and Private Information) and IMS-LIP [18,19] and they're the models that the next section presents, they organize a certain number of user data according to different structures.

1) PAPI Learner standard (Public and Private Information) aims to specify the semantics and syntax of learner's information (knowledge acquisitions, preferences, performance, skills, and relationships with other learners, etc.) [20]. In the PAPI model, a learner profile is defined by:

- PAPI Learner Performance: organize the information related to the learner's performance,
- PAPI Learner Personal: present personal information about the learner,
- PAPI Learner Preference: provide details Information related to learner preferences,
- PAPI Learner Portfolio: a collection of the learner's work "portfolio",
- PAPI Learner Relations: report relational information, relationships with other users (professors, other learners, etc.),
- PAPI Learner Security: register Security information like Passwords, keys, etc.

2) IMS LIP standard (Learner Information Package) is based on a data model that describes the characteristics required of a learner for general use. It defines a user data structure that can be imported or exported between interoperable systems [21]:

- Accessibility: Describes general accessibility (language skills, disabilities, eligibility requirements, and learning preferences),
- Activity: describes all relevant activities related to learning,
- Affiliation: gives information about the professional organizations to which the learner belongs,
- Competency: the learning skills acquired,
- Goal: assemble information about the learning goal, and other learner objectives such as personal goals and inspiration,
- Identification: learner's personal data (name, address, contact, …),
- Interest: information describing the learner's hobbies and recreational activities,
- QCL (Qualifications, Certifications & Licenses): lists the qualifications, certifications, and licenses obtained by the learner,
- Relationships: learner's relationship with other resources (identification, accessibility, activities, interests, etc.),
- Security key: security data (passwords, access rights, and security keys assigned to the learner),
- Transcript: presents a summary of the academic results.

IMS-LIP and PAPI Learner present a basic representation of the learner model. Yet IMS-LIP is a more stable extension of PAPI, with other specifications that are more detailed and more likely to contain other features [5, 22]. Without forgetting that the attributes of the two models are optional and reusable. Subsequently, to design a learner model for a particular situation or system, it is enough to have a certain number of predefined attributes, some of which may be optional, and to provide a framework to facilitate the creation of non-predefined attributes [23].

B. Covid-19 and Mental Health

The Quarantine, isolation, and social distancing have an impact on people's psychological well-being. Several studies have detected the harmful damage that the pandemic has on mental health: anxiety, fear, frustration, loneliness, anger, boredom, depression, stress, etc. [24].

Sanguino et al. found out that 18.7% of the people they diagnosed revealed depressive symptoms, 21.6% anxiety, and 15.8% Post-Traumatic Stress Disorder (PTSD) [25]. Under the same theme the results of Cao et al. were related specifically to students because they are among the groups of people that were most affected by the pandemic damages. They state that 24.9% of learners suffered from anxiety due to the COVID-19 epidemic and almost 1% of them suffered from severe anxiety [26].

Anxiety, depression, and PTSD are the most mental disorders found among students in times of crisis. Thus, the approach proposed will focus on adding these disorders to the learner model. The question that now arises is how to diagnose learners to find out if they are at risk of having a mental disorder. For this, screening tests and rating scales are useful in order to identify the presence of a disorder and to quantify its severity if it occurs [27].

1) Rating scales for depression: There are several scales and psychological tests that examine the presence and the severity of depression; the three golden standards are The Hamilton Rating Scale for Depression (HAM-D), The Beck

Depression Inventory (BDI), and Inventory of Depressive Symptomatology (IDS) [28].

The Hamilton depression scale (HAM-D) is among the most used scales to test depression clinically [29] with an optimized 17 item, it tests the existence and sorts the severity of depression: no depression (0–7); mild depression (8–16); moderate depression (17–23); to severe depression (≥24) [30]. The second one is The Beck Depression Inventory (BDI) which is a self-report scale designed to measure the severity of depressive symptoms; the first version contained 21 items associated to relative scores [31]. The third is the Quick Inventory of Depressive Symptomatology (QIDS), is a self-report version of the IDS 28 items that should be administered clinically, QIDS is a short questionnaire that takes 5 to 7 minutes with 16 items that reveal the severity of symptoms, and symptomatic change. The severity of depression ranges from 0 to 27 as follows (0-5) None, (6-10) Mild, (11-15) Moderate, (16-20) Severe, (21-27) Very severe [32].

2) Rating Scales for anxiety: Specialists can administrate criterions of general or specific anxiety disorder to get an idea of the level of anxiety of a patient [33]. There are several measures of anxiety, and in this section, we will focus on measures of general anxiety (GAD) such as The Hamilton Rating Scale for Anxiety (HAM-A) which is a 14-item anxiety symptom scale, administered by a clinician to measure the occurrence and severity of the disorder from 0 (not present) to 4 (very severe) [34]. And the Generalized Anxiety Disorder (GAD-7) scale, a 7-questions scale in its first version as the name suggests which was subsequently developed into a 13-questions questionnaire adding questions to assist the duration of symptoms [35]. GAD-7 has demonstrated to have good reliability to test for anxiety disorders and to sort out symptoms of mild, moderate, and severe anxiety [35-37].

3) Rating Scales for post-traumatic stress disorder (PTSD): A trauma is "the experience, witnessing, or confronting of an event that involves actual or threatened death or serious injury, or other threat to one's physical integrity" [38] and for the person to react with "intense fear, helplessness, or horror" [38], it exists plenty of scales to assess the PTSD symptoms like Clinician-Administered PTSD Scale (CAPS), an established interview of 10 items that assess the symptoms. The total severity score for the CAPS (CAPS-total) ranges between 0 and 136. Although it's time-consuming [39]. And the PTSD Checklist (PCL) which is a self-report scale that consists of 17 items to rate symptoms from 1 to 5 according to their severity [40], it helps to distinguish between patients at high or low risk of having an anxiety disorder [41].

III. Ontology Construction Methodology

Ontologies represent complex concepts and they have the possibility of being reusable and extensible and even interoperable with content on the web. The use of ontologies to model learners subsequently facilitates the adaptation of educational objects by adaptive systems. The learner model ontology proposed is structured according to the IMS Learner standard (IMS LIP), which allows to take the attribute needed, since all elements in LIP are optional and even add other attributes that will be pertinent to our context [14-13].

Our aim is to connect potential mental disorders with the relevant characteristics of the learner, in addition to data related to Covid-19 infection.

A. Process

In this work, we followed the Knowledge-Engineering Methodology which is an iterative process in 7 steps. It starts with determining the domain and scope of the ontology and leads us to create our instances [14].

1) Determine the domain and scope of the ontology: The first step is to define the scope of the ontology by deciding on which questions it should provide answers to (competency questions) [42]:

Q1: What are the essential characteristics of a learner?

Q2: What are the effects of the pandemic on the learner's mental health?

Q3: Is the learner or a family member was affected by the virus?

From these questions, we can see that the ontology will include various information about the learner both essential information out of the crisis and information about the learner's mental health and detect if he was affected by the pandemic or in the general time of crisis.

2) Consider reusing existing ontologies: In this step, we have studied the standard learner models and chosen the most used which are presented in section 2. We will reuse the attributes of the IMS-LIP standard because of its stability and its ability to exchange all types of information about the learner, so it can manage other functionalities required by new uses [22]. In addition to existing ontological modules such as Labib et al. who have worked on learning styles [4]. Zine et al. who regroup several standard models in order to realize its ontology [18]. Or [43] and [5] who have developed two ontologies in order to model lifelong learning with different relevant components. We have also taken into account the concept of reuse when developing our ontology, to ensure the possibility of reusing it in different systems, contexts and scenarios, for example, at different times of crisis such as natural disasters or even personal crises such as the death of a loved one, and not only in the case of a pandemic, which has inspired our research.

3) Enumerate important terms in the ontology: This step, as its name indicates, makes it possible to list the terms related to the ontology (terms that we would discuss about and their properties, the relationship between them, etc.). To create a useful list, we started by writing terms related to learner modelling, we scanned articles on learner modelling [3, 10, 12, 44, 45], and on mental disorders that are relevant to our research [27, 36, 38, 46] which we discussed in the second part. In addition to the various glossaries. Table I arranges some terms we have listed.

TABLE I. COLLECTION OF RELEVANT TERMS

Learner	Mental disorder	Pandemic
Knowledge	Anxiety	Covid-19
Level	depression	Confinement
competence	PTSD	isolation
characteristic	FEAR	virus
identity	symptoms	infected
skill	scale	transmission
performance	Standard assessment	Cluster site
degree	Severity	health crisis

4) Define the classes and the class hierarchy: We started by grouping similar items into classes using a combination of top-down and bottom-up approaches. First, we chose the most relevant and independent terms, then organize the specific information. We have tried to keep some attributes of the IMS-LIP in order to meet the standard. At the end of this step, we have obtained the following classes:

- Information: this class regroup the different IMS-LIP attributes.
- Mental disorder: represents the mental disorders that learners may have due to the pandemic and confinement.
- COVID-19 History: to determine if the learner or a family member is or has already been infected with the virus, has spent a period of confinement, lives in cluster site (Place where the number of cases is higher than expected).

Fig. 1 shows the different classes of our ontology.

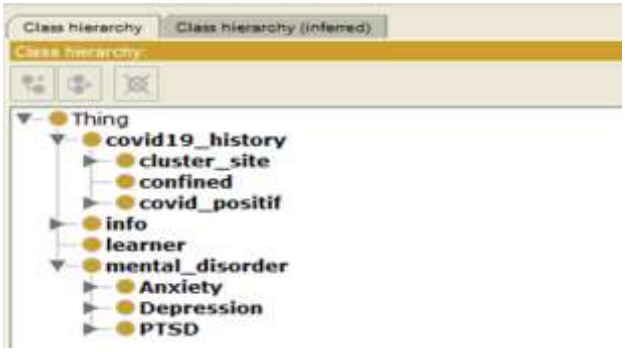


Fig. 1. The Upper-Level Concepts.

5) Define the properties of classes: After selecting classes, the next step is determining the properties/relationships to specify taxonomies for classes and properties. (e.g., "has_crisis"," has_covid_history,etc).

6) Define the facets of the slots: In this step we determine the slots cardinalities, type values, its range and domain, to limit the possible value for example "anxiety_level" is of type symbol and take the values (mild, moderate, severe).

7) Create instances: The last step is creating individual instances of classes in the hierarchy as well as the slot values for specific classes.

IV. ONTOLOGY IMPLEMENTATION

We chose the Web Ontology Language (OWL) of the World Wide Web Consortium's Web to encode our ontology and Protégé 4.3 ontologies editor to develop it. OWL helped us to specify the taxonomy for the classes and the different properties. Then we checked it with Protégé standard reasoner.

The proposed approach evolves from the standard learner model and existing ontologies cited in the previous section. However, it integrates other useful information about the COVID-19 pandemic and learner mental health and ensures a better representation of the learner.

We have organized the characteristics of our learner model into facets. The class of learner as shown in Fig. 2, is the key concept of our ontology, it includes all the details about the learner, and it is associated with other classes via has_info, has_crisis, and has_covid.

Info (Cf. Fig 3): this class answers the first question asked when establishing the scope of the ontology on the essential characteristics of the learner. Therefore, the class is composed of the different attribute of IMS LIP Identification as defined in the second section, Activity, Transcript, Interest, Competency, Accessibility, Security, and Affiliation. For example, as shown in the figure bellow the subclass QCL has degree which contains all the degrees of the learner (High School, Bachelor, Master, Doctoral) and his certification. The subclass competency class the competence of the learner in one of three levels {beginner, intermediate, expert}.

Mental_disorder (Cf. Fig. 4): The class presents relevant information about the learner's mental health after a crisis. Citing the troubles which might affect the learners. To test or diagnose the learner we choose some psychological scales as described in Section 2. This class answers the second question related to the effects of the pandemic and the confinement on the learner's mental health.

Covid_history (Cf. Fig. 5): The last important class is the one related to COVID-19 history (the last question of the first step). This class allow us to determine if a learner or a member of his family is affected by the virus, if he's confined in the moment and also if he lives in a cluster site (A specific site where the number of cases of an infectious disease that occurs over a specific period of time is higher than the expected number).[1]

[1] https://www.btb.termiumplus.gc.ca/publications/covid19-eng.html

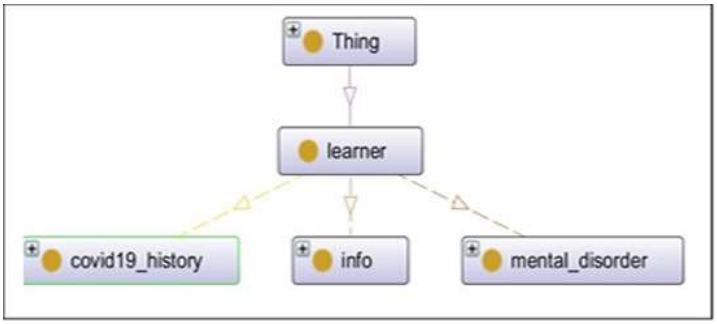


Fig. 2. The Learner Graphs.

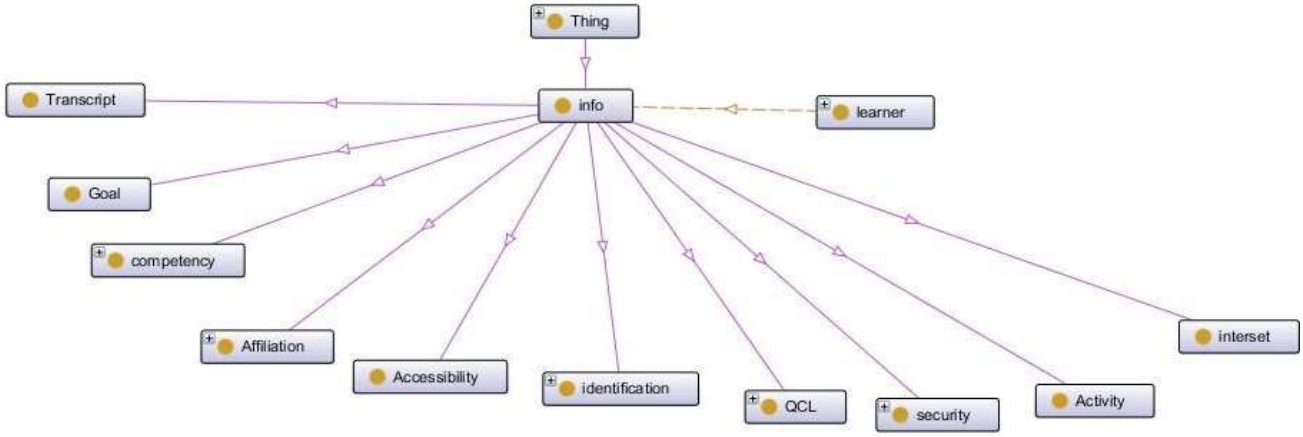


Fig. 3. The Info Class.

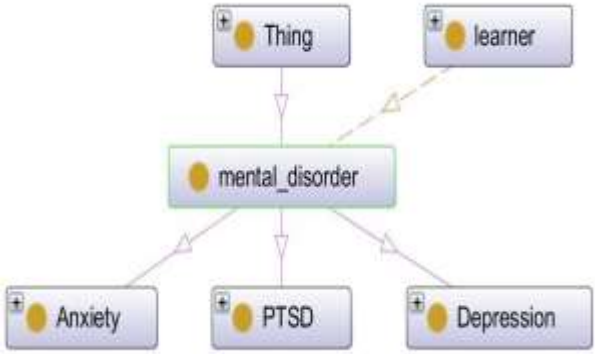


Fig. 4. Mental Disorder Class.

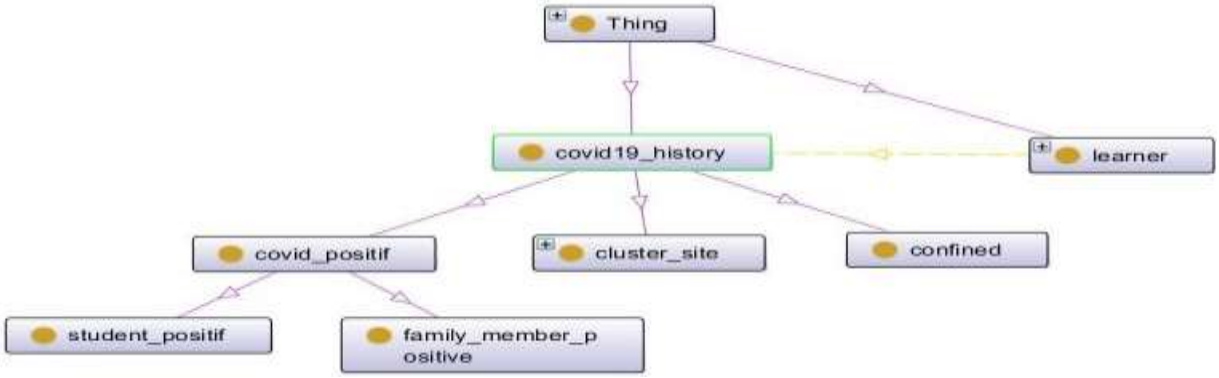


Fig. 5. Covid_History Class.

Fig. 2 to 5 illustrate the graphical representation of the proposed learner ontology. They capture all the concepts that describe learner profiles and are intended to answer questions about learner characteristics in times of crisis and beyond. Our ontology is IMS-LIP compliant.

Our main objective is to relate potential mental disorders to relevant learner characteristics, including data related to Covid-19 infection as shown in Fig. 6, which details the hierarchies of our ontology.

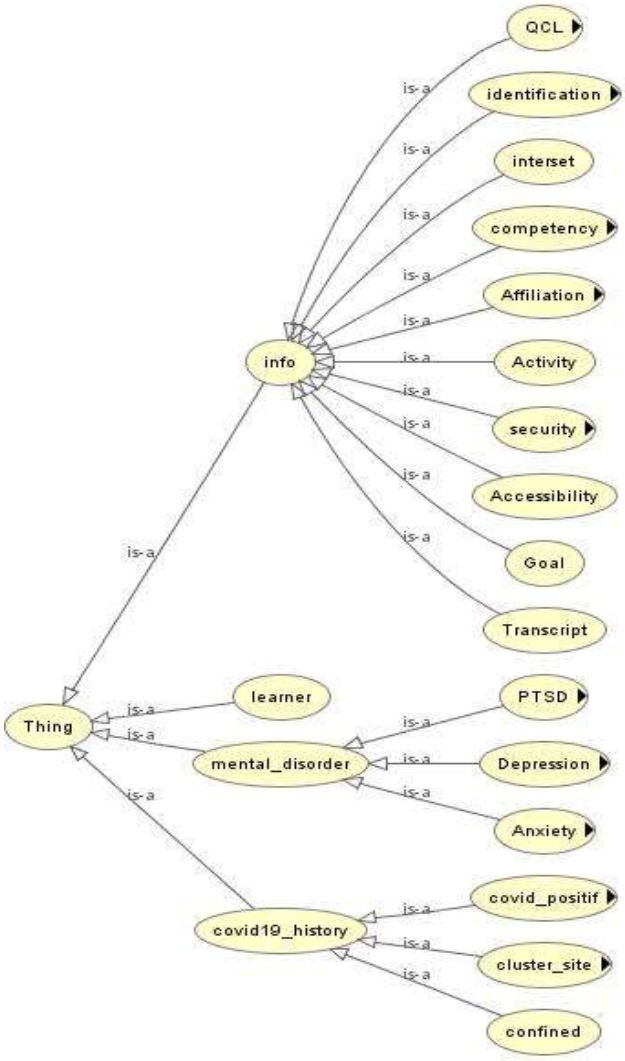


Fig. 6. Learner Model Ontology.

V. Discussion and Suggestion

There are several ontological models in the literature [4-7, 9, 18] presenting different aspects and characteristics of the learner such as abilities, learning style, prior knowledge, preferences, motivation, goals and many others. However, none of these models address criteria such as the mental disorders that the learner may have, especially in times of crisis, also the covid_history feature that is special to the time of pandemic and will be in use for the school and universities to control the pandemic situation.

Table II gives a comparison between the different learner models [5-8,17,18,43,] and the proposed model. It shows the difference between the models in the literature and our model according to the learner characteristics. The COVID-19 history data were not discussed due to its direct connection to the pandemic situation and never been discussed before in any work.

The table confirms that the mental health of learner is not taken in consideration despite its relevance to the learner ability to learn.

The proposed approach aims to give a representation that brings together the pedagogical and psychological characteristics, but also the health risks generated by the epidemic on the learners, by answering the three questions specified in the previous section. The first question with the objective of specifying the essential characteristics of the learner. Thus, our ontology defines the main characteristics of the learner (coordinates, diplomas, level, interest, etc.). It can be used in adaptive hypermedia systems, to propose specific courses according to the learner's objective and abilities, also according to his level of knowledge after having passed tests to detect it. The system will adapt a set of courses according to the learner's profile. During the learning session, the system can update learner information.

The second question addresses the damaging effects of the pandemic on the mental health of learners. According to Cao et al. anxiety, depression, and Post Traumatic Stress Disorder (PTSD) are the most diagnosed mental disorders among learners [26]. Our ontology can be used in systems to detect these disorders through psychological self-assessments that the learner can go through in order to provide psychological support or counselling, and eventually enable them to overcome these difficult circumstances. Or in educational systems to suggest less overwhelming educational objects. To give him more time for his homework, or other, taking into consideration his psychological state.

TABLE II. Comparative Study between the Proposed Model and Surveyed Model

Learner model ontology		**[5]**	**[6]**	**[7]**	**[8]**	**[17]**	**[18]**	**[43]**
Personal data	Accessibility				+		+	+
	Affiliation	+			+		+	
	Identification	+	+	+	+	+	+	+
	Interest	+	+		+		+	
	Security		+		+		+	
Pedagogical data	Activity		+		+		+	
	Competency	+	+	+	+	+	+	+
	Goal	+	+		+		+	+
	QCL	+	+		+		+	+
	Relationships	+			+			
	Transcript	+		+	+	+		+
Psychological Data	Anxiety				+			
	PTSD							
	Depression							

The final question that the ontology should answer is that of the health of the learner in relation to Covid-19. The proposed ontology ensures the detection if the learner or a family member is affected by the virus. It can be used to warn classmates, teachers and others who may be in contact with him, or in other examples if one of the learners' reports that he is living in a cluster site, this information can be used to identify others who are living in the same neighborhood to transfer them directly to distance learning.

The proposed ontological model is able to answer all three research (competency) questions. It is subsequently able to provide a representation of learners during and after the COVID-19 pandemic and in other times of crisis.

Reusability is also a very important part of our work. Besides the three examples of uses presented. The proposed ontology can be used and reused in different situations and different areas, not only in the case of a pandemic but in other times of social or personal crisis, including, but not limited to, natural disasters, terrorist attacks, loss of a loved one, divorce, etc. The proposed ontology allows the detection of the three mental disorders, but it gives the possibility to add others with their different scales according to the need. It gathers pedagogical, psychic, and health criteria related to COVID-19, but we can always add other characteristics on the learner such as motivation, commitment or other to enrich the ontology.

VI. Conclusion

In this paper, we modelled, created, and presented a learner model ontology during a time of crisis precisely in the Covid-19 pandemic. The ontology was constructed according to the knowledge engineering methodology using the Protege 4.3 Ontology Editor, and validated by the integrated HERMIT1.3.8 reasoner to validate its consistency, and to verify that it did not contain contradictory classes.

The focus was on the mental and physical health of the learners during and after Covid-19. The recorded information about the learner can be categorized into three main classes, the first one capturing academic information compliant with the IMS-LIP standard. The second collects information related to the mental disorders that the learner may have (anxiety, depression, PTSD) and their respective degrees of severity. The last one is the most related to Covid-19, it gathers the necessary information on the state of health of the learner if he is contaminated or at risk of contracting the virus.

Our approach gives the possibility to add other mental disorders with their scale for diagnosis if necessary and even other scale to the disorders that were mentioned in several conditions of a crisis in society (pandemic, natural disaster, terrorist attack, ...) or even in the case of personal or family crises (loss of a loved one, divorce ...). The choice to use ontologies to model our learner profile is as a consequence of their reusability, and in a context of adapting educational content, ontologies allow the semantic annotation of data and offer a better organization, indexing and management of data in order to provide the learner with relevant educational supports according to his profile. This ontology is a work in progress that we are working on improving considering different characteristic as the learning styles, the preferences, and the different constraints that the learner might be confronted with as low internet connection or lack of it.

The next step will be to integrate the ontology in an adaptive system, regroup similar learners for a better collaborative work, all to assure a good learning experience for every learner despite the circumstances.

Acknowledgment

This work is supported and financed by the University Hassan II of Casablanca- Morocco and The National Scientific research and technology Center (CNRST) under the program "support program for scientific and technological research related to covid-19", project "Elaboration of a psychological and pedagogical support platform".

References

[1] U. Nations and Secretary-General, "Policy Brief: Education during COVID-19 and beyond," 2020. [Online]. Available: https://unsdg.un.org/resources/policy-brief-education-during-covid-19-and-beyond.

[2] S. Cassidy, "Exploring individual differences as determining factors in student academic achievement in higher education," Studies in Higher Education, vol. 37, no. 7, pp. 793-810, 2012/11/01 2012, doi: 10.1080/03075079.2010.545948.

[3] A. Abyaa, M. K. Idrissi, and S. Bennani, "Learner modelling: systematic review of the literature from the last 5 years," Educational Technology Research and Development, vol. 67, no. 5, pp. 1105-1143, 2019.

[4] A. E. Labib, J. H. Canós, and M. C. Penadés, "On the way to learning style models integration: a Learner's Characteristics Ontology," Computers in Human Behavior, vol. 73, pp. 433-445, 2017.

[5] K. Rezgui, H. Mhiri, and K. Ghédira, "An Ontology-based Profile for Learner Representation in Learning Networks," International Journal of Emerging Technologies in Learning, vol. 9, no. 3, 2014.

[6] S. Ouf, M. Abd Ellatif, S. E. Salama, and Y. Helmy, "A proposed paradigm for smart learning environment based on semantic web," Computers in Human Behavior, vol. 72, pp. 796-818, 2017.

[7] S. A. Hosseini, A.-R. H. Tawil, H. Jahankhani, and M. Yarandi, "Towards an Ontological Learners' Modelling Approach for Personalised E-Learning," International Journal of Emerging Technologies in Learning, vol. 8, no. 2, 2013.

[8] D. Milosevic, M. Brkovic, and D. Bjekic, "Designing lesson content in adaptive learning environments," International Journal of Emerging Technologies in Learning (iJET), vol. 1, no. 2, 2006.

[9] R. Hammad and M. Odeh, "eLEM: A novel e-learner experience model," International Arab Journal of Information Technology, vol. 14, no. 4A, 2017.

[10] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," in The adaptive web: Springer, 2007, pp. 3-53.

[11] D. Vrandečić, "Ontology evaluation," in Handbook on ontologies: Springer, 2009, pp. 293-313.

[12] V. Vagale and L. Niedrite, "Learner Model's Utilization in the E-Learning Environments," in DB&Local Proceedings, 2012: Citeseer, pp. 162-174.

[13] S. Somyürek, "Student modeling: Recognizing the individual needs of users in e-learning environments," Journal of Human Sciences, vol. 6, no. 2, pp. 429-450, 2009.

[14] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," ed: Stanford knowledge systems laboratory technical report KSL-01-05 and …, 2001.

[15] M. Al-Yahya, R. George, and A. Alfaries, "Ontologies in E-learning: review of the literature," International Journal of Software Engineering and Its Applications, vol. 9, no. 2, pp. 67-84, 2015.

[16] M. Winter, C. A. Brooks, and J. E. Greer, "Towards Best Practices for Semantic Web Student Modelling," in AIED, 2005, pp. 694-701.

[17] A. Behaz and M. Djoudi, "Approche de Modélisation d'un Apprenant à base d'Ontologie pour un Hypermédia adaptatif Pédagogique," in CIIA, 2009.

[18] O. Zine, A. Derouich, and A. Talbi, "IMS Compliant Ontological Learner Model for Adaptive E-Learning Environments," International Journal of Emerging Technologies in Learning (iJET), vol. 14, no. 16, pp. 97-119, 2019.

[19] A. Paramythis and S. Loidl-Reisinger, "Adaptive learning environments and e-learning standards," in Second european conference on e-learning, 2003, vol. 1, no. 2003, pp. 369-379.

[20] IEEE P1484.2.1/D8, PAPI Learner — Core Features, I. C. Society, 2001. [Online]. Available: https://studylib.net/doc/18787880/ieee-p1484.2.1-d8--papi-learner-%E2%80%94-core-features.

[21] IMS Learner Information Packaging Information Model Specification v1.0, I. G. L. Consortium, 2001. [Online]. Available: http://www.imsglobal.org/profiles/lipinfo01.html.

[22] O. HULL, "Metadata standards for the description of portal users: a review," 2003.

[23] C. Jacquiot, "Modélisation logique et générique des systèmes d'hypermédias adaptatifs," PhD thesis, Department of Computer science, France, Supelec, 2006.

[24] D. Talevi et al., "Mental health outcomes of the CoViD-19 pandemic," Rivista di psichiatria, vol. 55, no. 3, pp. 137-144, 2020.

[25] C. González-Sanguino et al., "Mental health consequences during the initial stage of the 2020 Coronavirus pandemic (COVID-19) in Spain," Brain, Behavior, and Immunity, vol. 87, pp. 172-176, 2020/07/01/ 2020, doi: https://doi.org/10.1016/j.bbi.2020.05.040.

[26] W. Cao et al., "The psychological impact of the COVID-19 epidemic on college students in China," Psychiatry research, p. 112934, 2020.

[27] M. Blais and L. Baer, "Understanding rating scales and assessment instruments," in Handbook of clinical rating scales and assessment in psychiatry and mental health: Springer, 2009, pp. 1-6.

[28] C. Cusin, H. Yang, A. Yeung, and M. Fava, "Rating scales for depression," in Handbook of clinical rating scales and assessment in psychiatry and mental health: Springer, 2009, pp. 7-35.

[29] J. B. Williams, "Standardizing the Hamilton Depression Rating Scale: past, present, and future," European Archives of Psychiatry and Clinical Neuroscience, vol. 251, no. 2, pp. 6-12, 2001.

[30] M. Zimmerman, J. H. Martinez, D. Young, I. Chelminski, and K. Dalrymple, "Severity classification on the Hamilton depression rating scale," Journal of affective disorders, vol. 150, no. 2, pp. 384-388, 2013.

[31] A. T. Beck, R. A. Steer, and G. Brown, "Beck depression inventory–II," Psychological Assessment, 1996.

[32] A. J. Rush et al., "The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression," Biological psychiatry, vol. 54, no. 5, pp. 573-583, 2003.

[33] L. Marques, A. Chosak, N. M. Simon, D.-M. Phan, S. Wilhelm, and M. Pollack, "Rating scales for anxiety disorders," in Handbook of clinical rating scales and as-sessment in psychiatry and mental health: Springer, 2009, pp. 37-72.

[34] E. Thompson, "Hamilton rating scale for anxiety (HAM-A)," Occupational Medi-cine, vol. 65, no. 7, p. 601, 2015.

[35] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for as-sessing generalized anxiety disorder: the GAD-7," Archives of internal medicine, vol. 166, no. 10, pp. 1092-1097, 2006.

[36] K. Kroenke, R. L. Spitzer, J. B. Williams, P. O. Monahan, and B. Löwe, "Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection," Annals of internal medicine, vol. 146, no. 5, pp. 317-325, 2007.

[37] E. Moreno et al., "Factorial invariance of a computerized version of the GAD-7 across various demographic groups and over time in primary care patients," Journal of affective disorders, vol. 252, pp. 114-121, 2019.

[38] A. P. Association, Diagnostic criteria from dsM-iV-tr. American Psychiatric Pub, 2000.

[39] D. D. Blake et al., "The development of a clinician-administered PTSD scale," Journal of traumatic stress, vol. 8, no. 1, pp. 75-90, 1995.

[40] F. W. Weathers, B. T. Litz, D. S. Herman, J. A. Huska, and T. M. Keane, "The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility," in annual convention of the international society for traumatic stress studies, San Antonio, TX, 1993, vol. 462: San Antonio, TX.

[41] A. J. Lang and M. B. Stein, "An abbreviated PTSD checklist for use as a screening instrument in primary care," Behaviour research and therapy, vol. 43, no. 5, pp. 585-594, 2005.

[42] M. Grüninger and M. S. Fox, "Methodology for the design and evaluation of ontologies," 1995.

[43] D. Nurjanah, "LifeOn, a ubiquitous lifelong learner model ontology supporting adaptive learning," in 2018 IEEE Global Engineering Education Conference (EDUCON), 2018: IEEE, pp. 866-871.

[44] M. R. F. Sani, N. Mohammadian, and M. Hoseini, "Ontological learner modeling," Procedia-Social and Behavioral Sciences, vol. 46, pp. 5238-5243, 2012.

[45] L. Oubahssi and M. Grandbastien, "From learner information packages to student models: Which continuum?," in International Conference on Intelligent Tutoring Systems, 2006: Springer, pp. 288-297.

[46] M. G. Craske, S. L. Rauch, R. Ursano, J. Prenoveau, D. S. Pine, and R. E. Zin-barg, "What is an anxiety disorder?," Focus, vol. 9, no. 3, pp. 369-388, 2011.

A Survey on Dental Imaging for Building Classifier to Benefit the Dental Implant Practitioners

Shashikala J[1]
Research Scholar, Department of Electronics and Communication, JAIN (Deemed-to-be University), Assistant Professor, BMS Institute of Technology & Management, Bengaluru, Karnataka, India

Thangadurai N[2]
Professor and Research Coordinator, Department of Electronics and Communication, Faculty of Engineering and Technology, Centre for Emerging Technologies, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India

***Abstract*—Endo-osseous implants are considered an ideal dental fixture. It is becoming the preferred choice of the edentulous patient to rehabilitate toothlessness because of their aesthetic and functional outcome. Despite the successful surgery and implant placement, complications occur, which may be related to several factors, like operative assessment, treatment planning, patient-related factors, surgical procedures, and surgeons' experience. Comprehensive radiological assessment plays a vital role in clinical analysis for better treatment planning, avoiding complications, and increasing the Implant's success rate. However, despite the variety of dental imaging, choosing the right imaging technology has become difficult for clinical experts. The investigative survey conducted in this paper aims to determine the correlation between different imaging modalities, their essential role in implant therapy. This review extensively discussed which types of computational operations applied to image modalities in the existing literature address various noises and other relevant issues. These study findings reveal significant issues with various dental imaging modalities and provide an understanding to bridge all existing research gaps towards building cost-effective classification and predictive models for accurate dental treatment planning and higher implant success rates.**

***Keywords*—Dental implant; complication; implant failure; dental imaging; pre-processing**

I. Introduction

It is said that the mouth is a mirror of health that reflects the health condition of a person, or in other words, it is a cautionary system for disease. The mouth consists of both teeth and gums; their health condition is significant for oral health as poor oral health leads to various fatal diseases, too [1]. Apart from the fatal diseases, an issue of Edentulism-(toothlessness) is found in both kids, adults, and old aged people due to respective reasons leads to the inconvenience of chewing the food so, poor nutritional intake and as a result a poor health condition [2]. However, the stage of Edentulism also creates an issue of the hollowness of speaking-(pronunciation) along with other discomforts [3]. The traditional treatments of bridging and dentures were adopted for a long time as this was only a choice of treatment for the condition of Edentulism. However, the modern technique, namely dental implant surgery, is gaining popularity as an alternative solution to meet the deficiency of natural teeth by artificial tooth replacement [4].The dental implant procedure is based on the conception of direct contact between bone and the metal implant-(osseointegration), whose ultimate objective is to restore all the functional and aesthetic aspects [5]. An extensive planning and clinical examination performed by the dentist before surgery as a minor causality may cause serious harm to the patient. Therefore, an implant's success depends on several mutual factors, like implant region, bone quality, medical history of patient, skills, and the surgeon or dentist's experience. One of the significant challenges in dental implantation practice is the complex surgical procedures, which require preoperative and postoperative evaluation for achieving a higher success rate in dental implants [6]. The preoperative evaluation includes various factors such as the patient's general health conditions, bone quality, alveolar bone axis, and transplant site. The postoperative evaluation is carried out after the implantation to prevent any bias and risk of failure. Medical imaging technology plays a crucial role in the preoperative evaluation process. It provides the patient's anatomical details for the dental Implant-based on the maxillofacial structure and the two-dimensional geometric projection, helping clinical experts decide whether the implant surgery is suitable for the patient [7]. A systematic radiographic evaluation can provide an effective direction for precise positioning, which has important clinical significance in terms of accuracy and functional and aesthetic effects of the Implant [8]. Many imaging techniques are used in clinical dentistry practices, including conventional radiographic images and Computed Tomography (CT) for preoperative assessment and analysis of the complex jawbone structures. However, each imaging modality has some advantages and limitations too. Therefore, choosing the most suitable imaging method for dental implants is still tricky in dental practices. Another major issue is that the dental imaging is mostly associated with the poor image quality and superimposition factors that need to be process with an effective image enhancement and pre-processing techniques. In order to make the dental image representation more explanatory, several studies on dental image analysis have been conducted using digital image pre-processing methods. The proposed study aims to determine the prevalence of digital imaging modalities in dental implants and how they can help improvise the dental implant success rate. Therefore, this paper conducts a review analysis to highlight the importance of various imaging modalities and pre-processing techniques to explore the research gap. The rest of the sections of this paper are organized as follows: Section II discusses the background highlighting complications in the implant procedure and dental implant failures. This section also

discusses how to improvise the dental implant success rate by prediction using image analysis. Section III presents a thorough analysis of what kinds of dental imaging are used and for what purposes. A comprehensive analysis is conducted on dental imaging modalities to highlight their importance and limitations in this section. Section IV presents an analysis of the current state-of-the-art, observing the trends towards adopting radiography and cone-beam computed tomography to avoid anatomical structures critical to dental implant surgery. Section V discussion and perspective are presented. In this, significant research direction based on evidential proofs, open research issues, and inferences is explored to develop effective predictive models to benefit dental implant practitioners. Finally, the overall contribution of this paper is concluded in Section VI.

II. Study Background

Success and failure are two critical terms in dental implantology. The term implant success can be an ideal clinical setting, meaning that the Implant is into the jawbone and functions well and pleasingly. The term implant failure refers to the loss of osseointegration. Another statement is that it is an initial instance at which the Implant's efficacy, evaluated quantitatively, drops below a cut-off value or specified level [9].Dental implants may fail for various reasons, with the scope that distinguishes between complications and implant failures. This study uses the term implant failure, which is the complete loss of osseointegration, and the severity of implants that require to be removed from the implant site. In order to avoid any form of ambiguity, the study made a distinction between discussing implant failure and complications. Implant complications can be stated as an event that requires quantifiable clinical attention, and if such measures are not taken, the outcome of the implant therapy may be impaired. Implant complications may be caused due to poor patient selection, inadequate pre-assessment of the patient. Also, the degree of complications that are difficult to control may lead to implant failure. Therefore, through the proper patient selection and treatment planning, surgical fixing of implants can provide long-lasting functional and aesthetic restoration to the Edentulous Patient. Various studies have attempted to identify and quantify the rate of dental implant-related complications. However, to date, no single standard system for classifying dental implant-related complications. The authors in [10] discussed specific categories of complications related to dental implants. Existing studies [11-14] suggested the classification of complications associated with implant therapy considering all factors and causes. Other studies [15-16] considered the classification based on the particular phase of implant treatment that they tend to occur. The work carried out in [11] and [6] performed a classification of complications based on surgical, bone loss, implant loss, peri-implant soft tissue mechanical factor, and aesthetic/phonetic factor. In [13], the authors discussed the classification of dental implant complications, mechanical, technical, and biological. Classification of Surgical complications, Biological complications, and Restorative complications is carried out in [14]. The existing work of [15][16] discusses surgical complication based on three factors viz. i) implant treatment associated (wrong angulation, the judgment of improper implant-site, and lack of Communication among dental disciplines), ii) anatomy associated (nerve injury, bleeding, Sinus membrane complication, and devitalization of adjacent teeth), and iii) procedure associated-(Mechanical complication), lack of stability, mandibular fracture, aspiration, and ingestion. The authors in [17] discussed reversible complications are obstructions that are either temporary or easily fixed.

A proper surgical procedure analysis, including careful radiograph analysis, is significant to reduce the possibility of any implant complications and dental implant failure. Closer evaluation of dental radiographs helps to establish an appropriate treatment strategy for implant patients [18]. Several reviews and remarks have been given since the past few decades that described the significance of imaging techniques in dental disciplines [19-20]. Dental imaging plays a major-role in implant procedures to determine comprehensive information about the patient's maxillo-facial area to understand whether the surgical procedure is suitable for the patient. However, the role of imaging is not limited to determining only the maxillo-facial area but also at different stages of the treatment processes, leading to the ease of surgical practice towards achieving higher success in dental implants [21-22].Imaging in dental treatment stage-1 subjected to patient diagnosis and clinical analysis conducted before implant surgery. Imaging evaluation assists the dentist in making a clinical decision and effective treatment planning based on past radiographs, medical history, and new radiographs evaluations that determine bone angulations, quality of bone, the critical structure of the maxillo-facials, presence of disease, and analysis of the implant site. In dental treatment phase-2, the role of imaging is to care about surgical intervention by assessing the surgical site and implant position during and after surgery and estimates the duration required for healing. Phase-3 of dental treatment begins after the intra-operative assessment and continues until the Implant remains in the jaw. At this stage, dental imaging helps determine the care plan. If any changes or complications are noted during this period, the necessary clinical steps are taken to prevent any possibility of the risk of failure.

However, despite the variety of dental imaging, choosing an appropriate imaging technique has become a challenging task for clinical experts. Each imaging modality is associated with certain advantages and limitations. One of the major issues encountered in the dental image is the poor image quality due to poor contrast, uneven illumination, low resolution, and noise inclusion during the dental image acquisition process. In order to avoid any ill-effect, the radiation is kept low while taking the dental X-ray. The dental x-ray constructed at low-radiation generates very poor-quality images with lower contrast and brightness, causing visibility differences during analysis. The specific noises during the radiography cause degradation to the dental image. [23]. Therefore, an effective mechanism should be implemented to enhance the quality of the image that can provide a significant clinical analysis in dental implant surgical procedures. The criteria that need to be considered as follows:

- The dental image must provide cross-sectional interpretations that describe the spatial relationships between internal structures.
- The dental radiograph should not be compromised with distortion to a greater extent. However, the smallest distortion can be considered with a predictable average error to obtain a quantified analysis and precise measurement.
- It should provide an accurate description of bone density and cortical plate thickness to achieve the initial equilibrium and stability in the Implant.
- Radiography must provide higher dimensional accuracy in implant treatment procedure that includes analysis of implant placement site, pre-existing pathological condition of the patient, and evaluation alveolar thickness.
- The imaging tool should be available/provided at a reasonable price, and radiology doses should be as little as possible.

At present, there is various research works carried out towards dental imaging. However, there is always an impediment towards accurate diagnosis when it comes to medical image processing, as it demands a higher degree of accuracy. Hence, the prime statement of the problem of the proposed study is "To explore the strength and effectiveness of existing methodologies associated with dental imaging approach with respect to classification." The next section discusses about the different dental imaging modalities highlighting their advantage and limitations.

III. Dental Imaging Modalities

In this modern era, a variety of imaging technologies are widely used in the dental field. The traditional implant practitioners depend on 2D radiography. The advancement in imaging technology provided a 3D imaging technique, which offers advanced clinical evaluation in dental implants [24] and [25]. This section presents the adoption of verities of imaging modalities and their uses in different dental implant disciplines.

A. Conventional Imaging Modalities

Two-dimensional conventional imaging aims to complement the clinical analysis in dental implants by gaining a deep understanding of the internal teeth structure and alveolar bone. The different conventional imaging modalities are illustrated as follows:

- *Periapical Radiography*-It offers a systematic detail about the anatomical structures like teeth and surrounding tissues around the implant site. It is used for preoperative assessment to understand the implant area's structure, vertical height, and bone quality. However, these imaging modalities may be difficult to adopt due to accurate instrument positioning support's unavailability. This imaging technique is associated with distortion and magnification, limiting the quantified bone quality assessment, and suffers from providing accurate spatial relationships between internal overlapping dental structures [26].
- *Cephalometric Radiography*- This helps to capture the image of the head with the mandible in a lateral view to examine the associations between teeth, jaw, and the remaining part of the facial skeleton. This technique outlines the geometrical structure of the anterior alveolar region. The limitation is that it only displays cross-sectional images of bones associated with low magnification and overlapping issues [27].
- *Panoramic Radiography*- It is an x-ray radiography image that captures the entire mouth structure in a single image representation using a tomographic technique. It visualizes both maxillary-(upper jaw) and mandibular-(lower jaw) dental curves and supporting structures. It is mostly adopted as an initial screening x-ray image to assess dental and bone support, identify affected teeth, and the condition of dental implants. This imaging technique is primarily used in the preoperative assessment to depict jaws in a single radiograph film or a charge-coupled device image receptor [28]. The distinct advantage of the panoramic imaging technique is that it offers a low patient radiation dose and is cost-effective in terms of time and computation complexity. It involves easy functioning and takes little time to capture the entire image of dentition in a single film or image receptor. Like other conventional radiography imaging techniques, it also has some limitations. Since this imaging technique is an extra oral technique, it does not provide delicate anatomy than periapical radiographs. It suffers from the issues like geometric distortion, superimposition, and magnification. Some other problems, like positioning error and technical/ processing error during panoramic radiography [29].
- *Digital Radiography*- It is direct digital radiography carried using several functional units that includes x-ray–sensitive plates, sensors, mechanism of dividing it into electronic segments, and transferred to a computer to present and store the image. Compared to conventional imaging modalities, direct digital radiography offers good image quality with very little radiation. Few studies have mentioned that the overall reduction of radiation dose is up to 80% [30] and about 50% to 70% radiation reduction in intra oral and extra oral digital imaging [31]. Direct digital radiography has reduced processing time; images can be obtained immediately during the surgical procedure. Since this image is stored and processed in a computer, it can be manipulated with software programs to obtain enhanced visualization and accurate measurement. However, one of the significant disadvantages of digital radiography techniques is that the localization of sensors in the implant site sometimes becomes very challenging due to sensor size and positioning of the connecting cord.

Various conventional imaging modalities are discussed above. The limitation of conventional technology is that it encounters the superimposition of overlapping structures. The overlapping structure is caused due to the depiction of three-dimensional maxillofacial structures onto a two-dimensional

image plane, which results in the loss of spatial information that complicates the identification of objects of interest. The next sub-section discusses the applications and advantages of advanced imaging modalities in dental treatment.

B. Advanced Imaging Modalities

The conventional imaging modalities provide evidence for routine dentistry practices. Advanced imaging mechanisms are needed to demonstrate more information, complex diagnostics, and dental implant treatment plans. Hence, several techniques have changed the diagnosis and treatment planning strategies of dentistry. Some advanced dental imaging modalities are given below:

- *Computerized Axial Tomography (CAT)* - This is a unique X-ray imaging mechanism named computed tomography (CT), displaying the detailed images of the patient's anatomy with hard-and-soft tissues of the maxillofacial region. The CT uses multiple X-rays to construct a two-dimensional maxillofacial region and is converted into a three-dimensional image through processing. CT can obtain multiple, cross-sectional image-(slices) and generate high-contrast resolution images without suffering from superimposition and noise issues [32]. CT scans used to determine the quality of bones and the arrangement of teeth that cannot be efficiently obtained by the periapical imaging technique. CT identifies the diseases and immediacy of critical structures where implants are placed with the differentiation of tissues for analysis. The limitation of CT radiographs is that it has higher radiation exposure, high scan cost, and may not provide a good view of the small fissure resulting in false-negative readings [33].

- *Magnetic Resonance Imaging (MRI)* - MRI includes radio waves and adopts hydrogen atoms 'behaviour within a large magnetic field to look at body regions and generates an MR image of the internal structure. MRI represents soft tissue differences with high contrast sensitivity, which makes it advantageous over CT imaging. MR images can distinguish minor alveolar ducts and the contours between cortical bone and cancellous bone, thus obtaining necessary information about the maximum implant length, angle, and stability [34]. The MRI in the dental implant procedure seems to be an effective mechanism for 3-D imaging as it avoids the radiation risk of CT imaging. The adoption of MRI depends on the specific use conditions for an accurate diagnosis. The MRI achieves a flexible acquisition plane without changing image quality and resolution. However, MRI is susceptible to artifacts, distortion, and signal loss due to high magnetic susceptibility materials, while dental amalgam has little effect [35].

- *Cone Beam Computed Tomography (CBCT)* – CBCT is a variation of conventional CT. The application of CBCT is mainly for carrying diagnosis and planning of surgery in dental implants. One scan can produce many images of the area-of-interest. CBCT involves the mechanism of a cone-shaped-X-ray-beam moving around the patient to produce a large number of 2D views of ROI, and it is then converted into a 3D view using a cone-beam algorithm. CBCT in dentistry offers a high-resolution representation of bone and teeth, giving a spatial relationship between the adjacent structures. CBCT is used to evaluate osseous disease and identify jaw bone infections and diseases that help perform risk-free surgery, i.e. complications (pain and swelling) [36]. CBCT includes fast scanning procedures associated with lower radiation dose, lower scan cost and DICOM compatibility and has reduced metal product interference than other methods [37-38]. The limitation of CBCT is that it has a limited contrast range, gives fewer details of internal soft tissues, and has a large noise factor and artifacts.

C. Primary Findings

All the imaging methods have a vital role in dentistry applications. The conventional 2D and advanced 3D radiographs provide necessary information for dental treatment and Implant, while a dental digital panoramic image can offer a clinical diagnosis of the jawbone. The significance of digital panoramic imaging is that it has a low radiation dose and shorter exposure time [39]. But intraoral imaging has issues like low image quality, variable magnification, and ghost images. The superposition of the upper cervical spine is the main limitation of panoramic X-ray photography [40], and osseointegration cannot be detected due to overlapping issues [41]. Hence, it is limited to preoperative diagnostic, leading to implant failure [42]. Hence, implantation surgery may compromise the health of nearby soft tissues and cells [43]. The use of CT and CBCT is described in [44-45] over 2D radiographs to assess complex structures like the maxillary sinus. However, the limitations of these imaging modalities are i) not available in many local hospitals due to higher cost and multi-disciplinary technical requirements. The researchers also informed that the patients were exposed to higher radiation doses when CT examination is done than of 2D digital imaging and CBCT examination. Some research works also compared CBCT and digital panoramic imaging to assess the bone height towards planning treatment in different dental implant phases [46] and revealed that digital panoramic is self-sufficient to describe the incisor area but lacks in the canine area. Also, [47] have performed a comparison of error estimation and found CBCT has a better result, which holds a low average preoperative assessment error in the maxillary area than the digital panoramic imaging technique.

IV. State-of-the-Art Reviews

This section presents a review study on the state-of-the-art in the context of digital radiographs adopted in dental implant surgery. Digital radiography is cost-effective and is used in dental radiography. The study (Choi et al. [48]) investigates the impact of enhancement over periapical radiographs by considering three pre-processing techniques for diagnostics. The outcome gives quality differences between the processed image and the input image. A work of (Hao et al. [49]) considered denoising CBCT dental images where improved non-local means filtering is applied [49]. The outcomes demonstrated in terms of PSNR and MSE. The segmentation

operation over digital radiograph image is performed in (Cunha et al.) for the accurate visualization of dental Implant and crestal bone line [50]. A contrast enhancement over Digitized film-based panoramic dental image using the CLAHE-Rayleigh is found in the study of (Suprijanto et al.). The study outcome shows that this method has achieved better performance in terms of PSNR [51]. The authors in the study of (Yin et al.) have used approaches of noise filtering technique for CBCT image based on thresholding mechanisms and wavelet transform [52]. (Mortaheb and Rezaeian) introduces an automated dental CT image approach for identifying the vertical structure and arrangement of the teeth [53]. The study (Lamecker et al.) focuses on automated segmentation operation for Computer-assisted craniomaxillo facial surgery using cone-beam volumetric tomography-(CBVT) dental image [54]. A noise that occurred by positioning error in the digital panoramic dental image is considered in the work of (Amiri and Moudi et al.) and (Kandan and Kumar), which achieves better visualization of the roots of maxillary teeth in the digital radiograph [55-56]. The work carried out by (Naik et al.) used the histogram equalization technique for enhancing the overall visualization of the digital radiographs for accurate analysis of the bone structure and quality [57]. The authors (Kamezawa et al.) used a multiple noise filtering approach for CBCT imaging for exposure radiation dose reduction in an automated guided patient positioning system [58]. An edge enhancement-based pre-processing technique is applied on panoramic X-Ray in the study (Jufriadif et al.) to detect proximal caries [59]. The work of (Supriyanti et al.) used a point processing mechanism for contract stretching of a digital panoramic dental image [60]. In the study of (Khatter et al.), the authors have applied a multi-scale retinex mechanism over CBCT to perform a precise assessment of root canal anatomy for endodontic therapy [61]. An image pre-processing I2I scheme based on neural network architecture is adopted in the research work of (Zhao et al.), which considers generative adversarial networks (GAN) to suppress ring artifacts [62]. Mean-shift algorithm-based image segmentation is adopted in the study of (Gunawan et al.). The authors have identified a fuzzy region in the segmented image and performed fuzzy merging processes based on similarity measurement [63]. A work towards brightness preserving in dental digital periapical images using entropy and histogram analysis is found in the study of (Qassim et al.) [64]. A most recent research work carried out by (Abdallah et al.) [65] have used Anisotropic filtering to eliminate noise, and Contrast Limiting Adaptive Histogram Equalization (CLAHE) to enhance contrast, and sharpness of the dental panoramic image.

V. Discussion and Perspective

Several radiographic modalities were described with their respective features and limitations. Each has its applicability in respective dental conditions to assist the dentists in planning, evaluation, and implant treatment. A precise strategy can reduce the surgical complexity and postoperative complications and lead to higher success considering both aesthetic and functional aspects. Therefore, suitable radiographic selection plays an important role, and the advanced 3D radiograph technique provides all the functional utilities compared to the conventional radiograph technique. Due to the cost factor, digital, panoramic radiography is in wide use. However, advanced imaging modalities like (MRI, CT, and CBCT) provide better visualization and compatibility with analysis tools so that many complementary and significant information for successful dental implant planning is made available. The MRI facilitates precise localization of the complex structures and useful when the differentiation of soft tissue analysis is requiring, but it carries artifacts like geometric distortion. CT imaging is more suitable for the analysis of bone quantity and quality because it can quickly cover the expanded anatomical area and generate images with reduced noise caused by the patient's movement. The advanced and recent modality, namely, Cone Beam CT (CBCT), offers fast data acquisition of the complete field of view with minimal radiation exposure. It is useful in the diagnosis and Endodontic treatment. In all the above discussed, dental imaging modalities suffer image quality degradation due to various factors like superimposition, geometric distortion, loss of signal, contrast, motion artifacts, and positioning errors that cause challenges during interpretation. The efficient pre-processing techniques can enhance image quality; thereby, significant interpretations for accurate treatment planning in the pre-assessment phase during surgery can be achieved. The post-surgery complication can be avoided to illuminate the possibilities of implant failure.

A systematic review of existing research literature with these imaging modalities is inferred, used while proposing models for segmentation of ROI and classification of complex anatomical structures of the oral region. This paper potentially identifies the trend of the pre-processing techniques adopted and also found that both 2-D dental radiographs and CBCT are advantageous over other modalities. It is recommended that adopting 2D dental imaging with an efficient pre-processing technique for enhancement will be a better choice in implant treatment planning and surgical process until CBCT matures. In the future, CBCT with efficient pre-processing for enhancement and noise filtering may provide a way better path towards an effective modality for successful dental implantation.

A. Research Gap

Based on the above discussion and review analysis, the significant open research problem is highlighted as follows:

No standard open-source dataset is available for the analysis of CBCT. In most research works, the dataset was either collected from the hospitals or considered based on the experimental setup. It has also been seen that few research works have considered dental image data from internet sources.

- Most image enhancement techniques are in the transform domain so that some artifacts may appear in the output image. As a result, it may lead to over-enhancement and issues related to the edge of the image.
- Lack of novelty is analyzed in most of the existing literature subjected to dental image pre-processing tasks. Most of the existing research works follow a similar pattern towards pre-processing the medical image. An improvement and optimization mechanism should be considered.

- The research works towards a predictive model have also not focused on the computational complexity associated with their prediction model for classification of the anatomical structure in preoperative assessment for the Dental Implant.
- Analysis of dental Images based on consideration of suitable parameters is missing in the existing literature. In order to perform effective image analysis, researchers must Analysis and evaluate image quality based on the HSV feature and statistics error metrics like Peak-Signal-to-noise-ratio, MSE-(Mean square error), SNR-(Signal to noise ratio), CNR-(Contrast to noise ratio), SD-(Spectral Distance) and SSIM-(Structural Similarity index.).
- Standard benchmarking is also missing in most of the existing image pre-processing methods.

VI. Conclusion

A dental implant is a complicated procedure that involves multi-disciplinary activities for treatment and surgical planning. Appropriate knowledge and understanding of the complexity and evaluation of implant failure factors is crucial for dental practitioners. Apart from this, digital imaging analysis is critical stage clinicians need to understand the technical parameters. However, equally, it is essential to manipulate these dental radiographs using a suitable pre-processing mechanism to know the potential factors associated with each stage of implant treatment. This paper has presented an investigative review analysis of different complications factors, various dental imaging modalities, and state-of-art pre-processing techniques. Finally, the proposed survey also explored the significant issues in the existing literature and discussed the significant point of highlighting the open research problem. Therefore, the proposed review works provide an effective future research direction for establishing predictive models with effective pre-processing schemes to benefit dental implant practitioners.

References

[1] Clark, Danielle, and Liran Levin. "In the dental implant era, why do we still bother saving teeth?." Dental Traumatology 35, no. 6 (2019): 368-375.

[2] Andersson, Lars, Jens O. Andreasen, Peter Day, Geoffrey Heithersay, Martin Trope, Anthony J. DiAngelis, David J. Kenny et al. "Guidelines for the Management of Traumatic Dental Injuries: 2. Avulsion of Permanent Teeth." Pediatric dentistry 37, no. 6 (2015).

[3] Emami, Elham & de Souza, Raphael &Kabawat, Marla &Feine, Jocelyne. (2013). The Impact of Edentulism on Oral and General Health. International journal of dentistry. 2013. 498305. 10.1155/2013/498305.

[4] Alajlan, Abdulrahman, AryafAlhoumaidan, AbeerEttesh, and Mazen Doumani. "Assessing Knowledge and Attitude of Dental Patients regarding the Use of Dental Implants: A Survey-Based Research." International journal of dentistry 2019 (2019).

[5] Oh, Ji-hyeon. "Recent advances in dental implants." Maxillofacial plastic and reconstructive surgery 39, no. 1 (2017): 33.

[6] Bryce, G., D. I. Bomfim, and G. S. Bassi. "Pre-and postoperative management of dental implant placement. Part 2: management of early-presenting complications." British dental journal 217, no. 4 (2014): 171.

[7] Gupta, Sarika, Neelkant Patil, Jitender Solanki, Ravinder Singh, and Sanjeev Laller. "Oral implant imaging: a review." The Malaysian journal of medical sciences: MJMS 22, no. 3 (2015): 7.

[8] Jayadevappa, Busnur Shilpa, G. S. Kodhandarama, and S. V. Santosh. "Imaging of dental implants." Journal of Oral Health Research 1, no. 2 (2010): 50-62.

[9] Esposito M, Hirsch JM, Lekholm U, Thomsen P. Biological factors contributing to failures of osseointegrated oral implants (I) success Criteria and Epidemiology. Eur J Oral Sci 1998;106:527-51;

[10] Hanif, Ayesha, Saima Qureshi, Zeeshan Sheikh, and Haroon Rashid. "Complications in implant dentistry." European journal of dentistry 11, no. 01 (2017): 135-140.

[11] Goodacre CJ, Kan JY, Rungcharassaeng K. Clinical complications of osseointegrated implants. J Prosthet Dent 1999;81(5):537-552.

[12] Goodacre CJ, Bernal G, Rungcharassaeng K, Kan JYK. Clinical complications with implants and implant prostheses. J Prosthet Dent 2003;90(2):121-132.

[13] Hanif, Ayesha, Saima Qureshi, Zeeshan Sheikh, and Haroon Rashid. "Complications in implant dentistry." European journal of dentistry 11, no. 01 (2017): 135-140.

[14] Guo, Q., R. Lalji, A. V. Le, R. B. Judge, D. Bailey, W. Thomson, and K. Escobar. "Survival rates and complication types for single implants provided at the Melbourne Dental School." Australian dental journal 60, no. 3 (2015): 353-361.

[15] Misch K, Wang H. Implant surgery complications: etiology and treatment. Implant Dent 2008;17(2):159-168.

[16] Misch, Kelly, and Hom-Lay Wang. "Implant surgery complications: etiology and treatment." Implant dentistry 17, no. 2 (2008): 159-168.

[17] (60) Park SH, Wang HL. Implant reversible complications: classification and treatment. Impl Dent 2005;14:211-220.

[18] Gupta, Sarika, Neelkant Patil, Jitender Solanki, Ravinder Singh, and Sanjeev Laller. "Oral implant imaging: a review." The Malaysian journal of medical sciences: MJMS 22, no. 3 (2015): 7.

[19] Vandenberghe, Bart. "The digital patient–Imaging science in dentistry." Journal of dentistry 74 (2018): S21-S26.

[20] Satpathy, Anurag, Rajeev Ranjan, SubhashreePriyadarsini, Somesh Gupta, Piyush Mathur, and Monalisa Mishra. "Diagnostic Imaging Techniques in Oral Diseases." In Medical Imaging Methods, pp. 59-95. Springer, Singapore, 2019.

[21] Vandenberghe, Bart. "The digital patient–Imaging science in dentistry." Journal of dentistry 74 (2018): S21-S26.

[22] Chandak, Shruti, Arjit Agarwal, Ashutosh Kumar, Rajul Rastogi, Pawan Joon, Asif M. Wani, and Yuktika Gupta. "Comparative Study of DENTA Scan and Radiography for Preoperative Assessment of Dental Implants." Annals of International Medical and Dental Research 4, no. 1 (2018): 26.

[23] Rahmi-Fajrin H, Puspita S, Riyadi S, Sofiani E. Dental radiography image enhancement for treatment evaluation through digital image processing. Journal of clinical and experimental dentistry. 2018 Jul;10(7):e629.

[24] White, Stuart C., and Michael J. Pharoah. "The evolution and application of dental maxillofacial imaging modalities." Dental Clinics of North America 52, no. 4 (2008): 689-705.

[25] Ríos-Santos, José V., Cristina Ridao-Sacie, Pedro Bullón, Ana Fernández-Palacín, and Juan J. Segura-Egea. "Assessment of periapical status: a comparative study using film-based periapical radiographs and digital panoramic images." Med Oral Patol Oral Cir Bucal 15, no. 6 (2010): e952-6.

[26] Tanwani, HemlataBhagwan, Sheetal Sameer Potnis, Sandesh Satish Baralay, and Sameer Sidagouda Patil. "Comparison of conventional and digital cephalometric analysis: A pilot study." Journal of Dental and Allied Sciences 3, no. 2 (2014): 80.

[27] Choi, Bo-Ram, Da-Hye Choi, Kyung-Hoe Huh, Won-Jin Yi, Min-Suk Heo, Soon-Chul Choi, Kwang-Hak Bae, and Sam-Sun Lee. "Clinical image quality evaluation for panoramic radiography in Korean dental clinics." Imaging science in dentistry 42, no. 3 (2012): 183-190.

[28] Jayachandran, Sadaksharam. "Digital imaging in dentistry: A review." Contemporary clinical dentistry 8, no. 2 (2017): 193.

[29] Langland OE, Langlais RP, Preece JW. Principles of dental imaging. 2nd ed. Philadelphia: Lippincott Williams & Wilkins, 2002: 285.

[30] Farman AG, Farman TT. Extraoral and panoramic systems. Dent Clin North Am 2000; 44: 257-272,

[31] Aggarwal V, Logani A, Shah N. The evaluation of computed tomography scans and ultrasounds in the differential diagnosis of periapical lesions.

[32] Shah, Naseem, Nikhil Bansal, and Ajay Logani. "Recent advances in imaging technologies in dentistry." World journal of radiology 6, no. 10 (2014): 794.

[33] Niraj, Lav Kumar, Basavaraj Patthi, Ashish Singla, Ritu Gupta, Irfan Ali, Kuldeep Dhama, Jishnu Krishna Kumar, and Monika Prasad. "MRI in dentistry-a future towards radiation free imaging–systematic review." Journal of clinical and diagnostic research: JCDR 10, no. 10 (2016): ZE14.

[34] Mendes, Silwan, Carin A. Rinne, Julia C. Schmidt, Dorothea Dagassan-Berndt, and Clemens Walter. "Evaluation of magnetic resonance imaging for diagnostic purposes in operative dentistry—a systematic review." Clinical Oral Investigations (2019): 1-11.

[35] Bornstein, Michael M., William C. Scarfe, Vida M. Vaughn, and Reinhilde Jacobs. "Cone beam computed tomography in implant dentistry: a systematic review focusing on guidelines, indications, and radiation dose risks." International journal of oral & maxillofacial implants 29 (2014).

[36] Gupta, Jyoti, and Syed Parveez Ali. "Cone beam computed tomography in oral implants." National journal of maxillofacial surgery 4, no. 1 (2013): 2.

[37] Bornstein, Michael M., Keith Horner, and Reinhilde Jacobs. "Use of cone beam computed tomography in implant dentistry: current concepts, indications and limitations for clinical practice and research." Periodontology 2000 73, no. 1 (2017): 51-72.

[38] Jacobs, Reinhilde, Benjamin Salmon, Marina Codari, Bassam Hassan, and Michael M. Bornstein. "Cone beam computed tomography in implant dentistry: recommendations for clinical use." BMC Oral Health 18, no. 1 (2018): 88.

[39] Suomalainen A, PakbaznejadEsmaeili E, Robinson S. Dentomaxillofacial imaging with panoramic views and cone beam CT. Insights Imaging 2015; 6: 1-16.

[40] Tang Z, Liu X, Chen K. Comparison of digital panoramicradiography versus cone beam computerized tomography for measuring alveolar bone. Head Face Med 2017; 13: 2.

[41] Isidor F. Clinical probing and radiographic assessment in relation to the histologic bone level at oral implants in monkeys. Clin Oral Implants Res 1997; 8: 255-64.

[42] Greenstein G, Cavallaro J, Romanos G, Tarnow D. Clinical recommendations for avoiding and managing surgical complications associated with implant dentistry: a review. J Periodontol 2008; 79: 1317-29.

[43] Tang Z, Liu X, Chen K. Comparison of digital panoramic radiography versus cone beam computerized tomography for measuring alveolar bone. Head Face Med 2017; 13: 2.

[44] Hassan B, Jacobs R. Cone beam computed tomography - 3D imaging in oral and maxillofacial surgery. Eur Med Imaging Rev 2008; 1: 38-40.

[45] Kopecka D, Simunek A, Streblov J, Slezak R, Capek L. Measurement of the interantral bone in implant dentistry using panoramic radiography and cone beam computed tomography: a human radiographic study. West Indian Med J 2014; 63: 503-9.

[46] Renton T, Dawood A, Shah A, Searson L, Yilmaz Z. Post-implant neuropathy of the trigeminal nerve. A case series. Br Dent J 2012; 212: E17.

[47] Angelopoulos C, Thomas S, Hechler S, Parissis N, Hlavacek M. Comparison between digital panoramic radiography and conebeam computed tomography for the identification of the mandibular canal as part of presurgical dental implant assessment. J Oral Maxillofac Surg 2008; 66: 2130-5.

[48] Choi, Jin-Woo, Won-Jeong Han, and Eun-Kyung Kim. "Image enhancement of digital periapical radiographs according to diagnostic tasks." Imaging science in dentistry 44, no. 1 (2014): 31-35.

[49] Hao, Jia, Li Zhang, Liang Li, and Kejun Kang. "An improved non-local means regularized iterative reconstruction method for low-dose dental CBCT." In 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), pp. 3422-3425. IEEE, 2012.

[50] Cunha, Pedro, Miguel A. Guevara, Ana Messias, Salomão Rocha, Rita Reis, and Pedro MG Nicolau. "A method for segmentation of dental implants and crestal bone." International journal of computer assisted radiology and surgery 8, no. 5 (2013): 711-721.

[51] Juliastuti, E., and LusiEpsilawati. "Image contrast enhancement for film-based dental panoramic radiography." In 2012 International Conference on System Engineering and Technology (ICSET), pp. 1-5. IEEE, 2012.

[52] Yin, Yong, Gang Yu, Hongjun Wang, Zhi Liu, and Dengwang Li. "CBCT image denoising based on multi-scale wavelet transform." In 2010 3rd International Conference on Biomedical Engineering and Informatics, vol. 1, pp. 150-153. IEEE, 2010.

[53] Mortaheb, Parinaz, Mehdi Rezaeian, and Hamid Soltanian-Zadeh. "Automatic dental CT image segmentation using mean shift algorithm." In 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP), pp. 121-126. IEEE, 2013.

[54] Lamecker, Hans, Stefan Zachow, Antonia Wittmers, Britta Weber, H. Hege, B. Isholtz, and Michael Stiller. "Automatic segmentation of mandibles in low-dose CT-data." International Journal of Computer Assisted Radiology and Surgery 1 (2006): 393.

[55] AsadiAmiri, Sekine, and Ehsan Moudi. "Image quality enhancement in digital panoramic radiograph." Journal of AI and Data Mining 2, no. 1 (2014): 1-6.

[56] Kandan, R. Somas, A. John, and S. Kumar. "An improved contrast enhancement approach for panoramic dental x-ray images." ARPN J Eng App Sci 10 (2015): 1897-1901.

[57] Naik, Anjali, Shubhangi Vinayak Tikhe, and S. D. Bhide. "Histogram Equalization for Class-Identification of Dental Disease Using Digital Radiography." In International Conference on Business Administration and Information Processing, pp. 144-151. Springer, Berlin, Heidelberg, 2010.

[58] Kamezawa, Hidemi, KatsutoshiShirieda, Hidetaka Arimura, Noboru Kameda, and Masafumi Ohki. "An approach of exposure dose reduction of cone-beam computed tomography in an image guided patient positioning system by using various noise suppression filters." In 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS), pp. 1475-1780. IEEE, 2014.

[59] Na'am J, Harlan J, Madenda S, Santony J, Suharinto C. Detection of proximal caries at the molar teeth using edge enhancement algorithm. International Journal of Electrical and Computer Engineering. 2018 Oct 1;8(5):3259.

[60] Supriyanti R, Setiadi AS, Ramadhani Y, Widodo HB. Point Processing Method for Improving Dental Radiology Image Quality. International Journal of Electrical and Computer Engineering (2088-8708). 2016 Aug 1;6(4).

[61] Khatter, Ashish, Anita Thakur, and Nitya Reddy. "CBCT Image Feature Enhancement for Endodontic Therapy." In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 293-296. IEEE, 2019.

[62] Zhao, Shuyang, Jianwu Li, and QirunHuo. "Removing ring artifacts in CBCT images via generative adversarial network." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055-1059. IEEE, 2018.

[63] Gunawan W, Arifin AZ, Indraswari R, Navastara DA. Fuzzy Region Merging Using Fuzzy Similarity Measurement on Image Segmentation. International Journal of Electrical & Computer Engineering (2088-8708). 2017 Dec 1;7(6).

[64] Qassim, Hassan M., Nasseer M. Basheer, and Mazin N. Farhan. "Brightness preserving enhancement for dental digital X-ray images based on entropy and histogram analysis." J Appl Sci Eng 22 (2019): 187-94.

[65] Yousif Mohamed, Nouf H Abuhadi, and Maryam Hasan Hugri"Enhancement of Dental X-rays Images Using Image Processing Techniques"Journal of Research in Medical and Dental Science 2021, Volume 9, Issue 2, Page No: 12-16.

Emerging Line of Research Approach in Precision Agriculture: An Insight Study

Vanishree K[1]
Research Scholar, Department of ISE
RV College of Engineering
Bengaluru, India

Nagaraja G S[2]
Professor and Associate Dean, Department of Computer Science and Engineering, RV College of Engineering
Bengaluru, India

***Abstract*—The present state of agriculture and its demand is very much different than what it used to be two decades back. Hence, Precision Agriculture (PA) is more in demand to address this challenging demand. With consistent pressure to develop multiple products over the same agricultural land, farmers find PA's adoption the best rescue-based solution with restricted resources. PA accelerates the yield and potentially assists in catering up the demand of scarcity of demands of food. With the increasing adoption of PA-based technologies over farming, there are best possibilities to explore efficient farming practices and better decision-making facilitated by real-time data availability. There isan evolution of various novel technologies to boost agricultural performance, i.e. variable rate technology, Geomapping, remote sensing, automated steering system, and satellite positioning system. Apart from this, it is also observed that Internet-of-Things (IoT) and Wireless Sensor Network (WSN) have been slowly penetrating this area to accelerate PA's technological advancement. It is noticed that the adoption of sensing technology is a common factor in almost all the techniques used in PA. However, there is no clear idea about the most dominant approach in this regard. Therefore, this paper discusses existing approaches concerning standard conventional PA and sensing-based PA using WSN. The study contributes towards some impressive learning outcomes to state that WSN and IoT are extensive to boost PA.**

Keywords*—Precision agriculture; smart farming; wireless sensor network; internet-of-things; remote sensing; variable rate technology

I. Introduction

Technological advancement has penetrated agriculture in the present time, right from small to large scale farming [1]. Two decades back, the Global Positioning System (GPS) usage permits the farmers to collect necessary farming data, which facilitates autonomous steering control system development [2]. However, the present times make use of more advanced technologies, e.g., fixed solutions for Internet-of-Things (IoT), aerial devices, sensors, etc., to carve the progressive path of Precision Agriculture (PA). The prime goal of PA is to achieve, i) opt for the appropriate crop to ensure increased quality yield and make more revenue in the commercial market, ii) using the proper data to assess the performance of the farming land, iii) improve the economics of farming and another offer better sustainability towards the environment, and iv) making a prediction of climatic fluctuations and taking necessary countermeasures to protect from upcoming threat towards agriculture [3]-[5]. The significant beneficial aspectof PA is minimizing and controlling crop waste and adverse influence over the environment.Farmers are facilitated with the appropriate anticipated yield for their farming land. Investigation towards PA could offer potential insight towards solving the crisis of food demand globally [6]. Farmers are now able to identify the beneficial aspects of PA introduced by IoT. The return of investment and quality of decision-making can be ensured by adoption PA by business owners.There is the inclusion of various metrics to carry out PA, e.g., fertilizer input, a sample of soil, nutrient availability of soil, rainfall level, temperature, etc. [7].

Acquisition of this information via sensors can lead to precision decision-making by the farmers. It can also furnish various real-time data of their farming land, identifying specific production patterns or identifying any associated risk factors during cultivation and harvesting season. Adopting PA also facilitates exclusive access to the agricultural records via cloud-based resources where the data can be accessed anytime and anywhere [8]. It also leads to an adequate formulation of measures towards crop protection. Usage of sensors can quickly identify the health statistics of a plant concerning soil pressure, presence of chemicals, environmental impact, pest, etc. [9]. This information leads to a better decision in planning for fertilizer input by the farmer. The most potential benefit of PA is associated with irrigation management. Any form of the crop demands an adequate water supply in appropriate quantities and channel them throughout the farming land. Usage of various controllers, actuators, and sensors further offers relevantwater supply statistics for better irrigation management. To effectively operational, PA demands the use of progressive technologies, i.e., usage of sensors [10], precision farming software [11], connectivity protocols [12], and location monitoring tools [13]. Irrespective of PA's known benefits, it is still yet to get a discloser about the research progress regarding more insights over challenging state of farming, minimal resource waste, identifying the unique pattern of production or risk. Therefore, this manuscript offers an exhaustive review of standard and upcoming potential PA approaches to providea more precise research state. The significant contributionsin the proposed paper aredescribed as follows:

- The present state of conventional approaches in PA is highly scattered. So this paper contributes towards offering a compact discussion of conventional standard approaches concerning its taxonomies.

- Presents an elaborative discussion of all the potential implementation carried out in present times towards conventional standard approaches in PA.
- Discussion about the existing approaches carried out by Wireless Sensor Network (WSN) to identify the strength and weaknesses.
- Presents a compact discussion about existing research trends to have a real picture of existing approaches, targeted issues, and technological adoption.
- Contributes towards more in-depth insight of the study findings concerning learning outcomes to visualize the clear picture of PA approaches

The remaining sections of the proposed manuscript areorganized as follows: Section II discussesthe essentials of precision agricultureconcerning all standard taxonomies and conventional research-based approaches. Since WSN is identified as upcoming technology and IoT in precision agriculture, Section III discusses various techniques used in WSN in precision agriculture; Section IV discusses the research trend. In contrast, Section V highlights about learning outcome of this manuscript. Finally, Section VI summarizes the overall contribution of the proposed review study and briefs about future work direction in precision agriculturebased on study findings.

II. Precision Agriculture

Precision Agriculture (PA) targets improving crop production with the adoption of advanced technologies. This concept deals with improving agricultural management based on various scientific observations [14]. The primary aim of precision agriculture is to construct an appropriate decision-making system capable of optimizing productivity without consuming expensive resources [15]. It is believed that crop production is significantly affected by the terrain features studied in the phytogeomorphological mechanism [16]. The evolution of the phytogeomorphological mechanism is due to the realizationthat the hydrological factors of farmland are controlled by geomorphic components [17].The proliferation of various satellite navigation systems has further boosted the adoption of precision agriculture [18]. Adopting such a navigation system helps localize an appropriate location of the agricultural land suitable for production. Such geographic information obtained from satellite navigational system also furnishes spatial information of land concerningactual contents required for cultivation viz. potassium, manganese, pH level, nitrogen level, moisture level, crop yield, etc. [19]. A sensory-based satellite navigation system helps further more data collection, right from a degree of water in the soil to the level of chlorophyll. More granularity can be obtained from hyperspectral imaging in this regard. At present, there are different forms of variable rate technology (e.g., sprayers, seeders) that are used along with satellite images for optimizing the resources [20]. However, the current advances in technologies are more inclined to use sensors planted within the soil. This sensor can directly forward the aggregated data to the user autonomously without any dependency on human interactivity.

The adoption of airborne vehicles is also used in precision agriculture due to their cost-effective nature and does not require specialized skills to make them airborne. Such approaches make use of photogrammetric techniques by using different forms of the camera (for both color and hyperspectral images) are used over airborne vehicles to extract information associated with the field images [21]. The images obtained by this technique can be used for evaluating the different forms of vegetative index [22]. Apart from this, a different form of other information, e.g.,the elevation of land, can also be captured by airborne vehicles subjected to variousconditions of sophisticated software models for constructing topography [23]. Therefore, a better probability of enhancing crop cultivation can be achieved by studying such a topography map. This information can be used for improving the inputs towards healthy cultivation, e.g., growth regulators, different types of chemicals, fertilizers, water, etc. Therefore, using different forms of technologies in precision agriculture is used to study crop science, accelerate the economics associated with the production, and protect the environment by controlling different possibilities of risk and agricultural footprints.

A. Standard Taxonomies of Technologies in PA

The novel approaches of agricultural practices are now facilitated by the advent of different technologies in PA. The optimization is now possible for PA for both profitability and productivity based on decision-making and real-time information over the field. The prime targets of the technologies used in PA are mainly to control the agricultural input along with environmental protection. On this basis, it is seen that there are five standard taxonomies of precision farming,including 1) Satellite Positioning System, 2) Variable Rate Technology, 3) Geomapping, 4) Automated Steering System, and 5) Remote Sensing as in (Fig. 1).

In *Satellite Positioning System*, the prime technological contributor is the Global Positioning System (GPS), mainly using data associated with geo-references of production and auto-steer system. The agricultural machines (e.g., tractors) are better controlled with accuracy using GPS inbuilt within the machine. The farming operation is improved when the driver is provided with error-free information with machine movement patterns (Fig. 2).

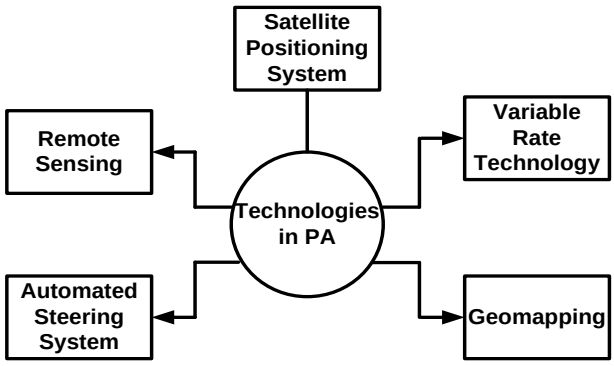
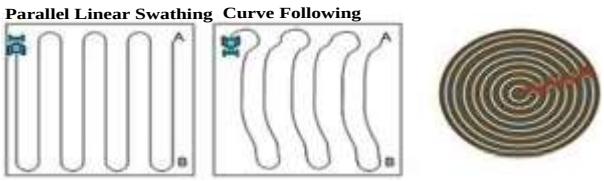


Fig. 1. Standard Taxonomies of Technology used in PA.



(a) (b) (c)

Fig. 2. Patterns of Field Traffic enabled by GPS. a) Linear-Parallel Pattern, b) Curve Pattern, and c) Circular Pattern.

In Variable Rate Technology, the agricultural inputs are controlled by farmers. Adopting this standard technology offers planting density to be optimized while increasing the applicate rate's efficiency towards nutrients and pest protection. This significantly minimizes the farming cost as well as effectively control the adverse impact onthe environment. When variable rate technology is integrated with application equipment, the system offers precise information about the field's location and appropriate time for obtaining input for rates corresponding to the region-specific application. Fig. 3 highlights the soil map used for variable-rate technology to find the different nutrients needed in the soil.

In Geomapping and Remote Sensing, sensors are usually used to construct a map with the different crop and soil conditions, e.g., pest, soil pH, type of soil, nutrient level of the soil. Sensors are attached to different machines and vehicles to be dominantly used for creating soil maps. Sensors collect the information from the field and GPS to assess the health statistics of crops and soil. This information is then passed on to a specific location in an area. Farmers can carry out identification of specific events or any significant alteration in the properties of soil. Fig. 4 highlights the mapped field which is used by the sensors built over the agriculture machine.

In the Automated Steering System, the vehicles used in agriculture are involuntarily steered by the navigation system. This technology reduces human-related errors while controlling the movement of the vehicle. It also permits effective management of the field by providing overhead tuning andcontrolling the machinerybased on edge information. The existing system uses differential correction for real-time kinematics to offer accuracy in the form of centimeters. Fig. 5 highlights overlapping factors of auto-steering system and manual machine.

However, to offer higher accuracy for the machinery over the deployed path, installing a specific communication system with a base station is required. A precise point positioning system does not require any form of data communication in the auto-steering system [24]. On the other hand, machinery can also be allowed to be moved using GPS based navigation system.

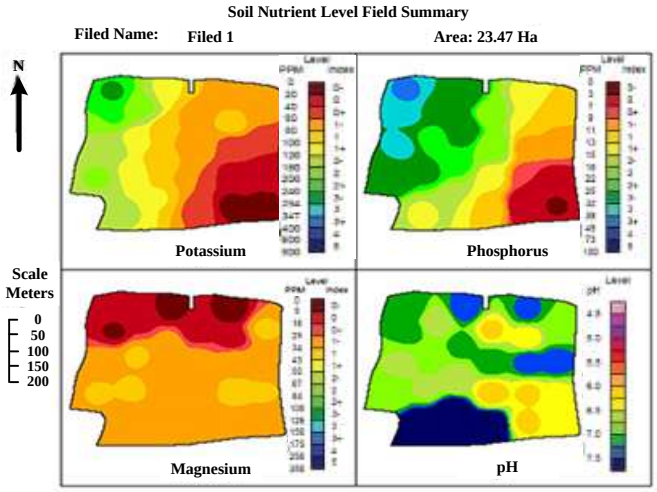
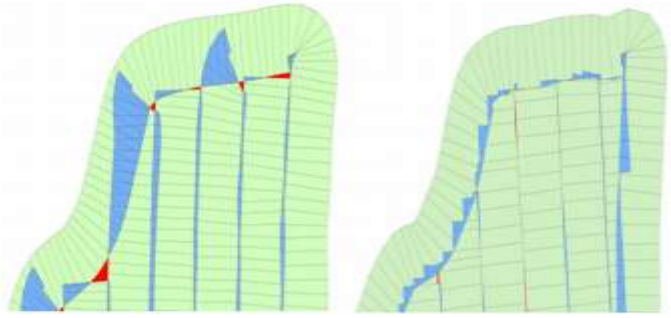


Fig. 3. Usage of Soil Map for Analyzing the Level of Nutrients in the Soil, a) Presence of Potassium, b) the Presence of Phosphorus, c) Presence of Magnesium, and d) the Presence of pH.

Fig. 4. Geomapping and Remotely Sensed Soil Map with its Properties from the Sensor Fitted in the Machine.



(a) (b)

Fig. 5. Automated Steering System. (a) Manual Machine Guided Field Overlapping (Blue), b) Auto-Steering based Field Overlapping.

B. Review of Studies on Conventional PA Technologies

This section discusses the various research work being carried out towards different standard technologies in PA briefed in the prior section.

- *Satellite Positioning System*: This approach uses two prominent techniques, i.e., GPS (Global Positioning System) and GNSS (Global Navigation Satellite System). It is found that GPS, when integrated with the robotic application, could significantly contribute towards PA. However, the GPS signal's availability could be impacted due to occlusion towards GPS-enabled Real-Time Kinematic (RTK) in farming. This problem is addressed in Levoir et al. [25] by evolving out with a smart rover that uses sophisticated image processing and statistical analysis to perform localization tasks by the rover. Further studies show that integrating GPS with the sensory application could improve the data acquisition with more accuracy (Rodriguez et al. [26]). A prototype was developed for herbicide ballistic technology integrated with sensors and GPS to automate data acquisition. Prototyping-based modeling is evolving in an existing system where GPS is integrated with a micro-electromechanical system.The idea was to offer a precise steering angle of the agriculture vehicle (Si et al. [27]). An unscented Kalman filter did the computation of the steering angle. Existing study towards the adoption of GPS has mainly emphasized achieving better accuracy for the receivers (Dabove et al. [28]).It should be noted that GPS is an integral section of GNSS with variable ranges of transmission frequency. Literature has also studied the adoption of GNSS towards precision farming (Marucci et al. [29]); however, it does not work effectively in hilly regions. There is still a better possibility of improvement when the GNSS system is combined with different technologies to overcome this issue. GNSS is also found with various artifacts, e.g., multipath error, atmospheric interference, satellite configuration (Stombaugh et al. [30]).

- *Variable Rate Technology:* This kind of technology is used for managing crop production specific to the farming region (Rubio and Mas [31], Ayaz et al. [32]). The recent work carried out by Nordblom et al. [33] have used variable rate technology in PA focusing on nitrogen fertilizer input. The study integrates such application with Geographic Information System (GIS) and rainfall data to determine the reason for waterlogging in a specific geographic area. The study has also simulated data distribution of financial risk in predictive mode to signify variable rate technology. A similar direction of work is also carried out by Steffani [34], where a statistical model is used for analyzing lint. The idea is to emphasize adequate control over the environment and maximization of profit, as discussed in the study of Kweon et al. [35]. A study carried out by Colaco [36] has analyzed the impact of this technology on yield, the fertility of the soil, and fertilizer consumption. The study outcome shows that the variability factor can successfully achieve increased production without much dependency on excessive fertilizers. A study carried out by Nawar et al. [37] highlights that this technology, when integrated with region delineation management approach then it could lead to better efficiency in farming in contrast to application with uniform rate. At present, the implementation of variable rate technology is further boosted by the proliferation of novel solutions by manufacturers of farming equipment. The work carried out by Thomasson et al. [38] has discussed the frequently adopted manufacturers using crop sensors associated with this technology of nitrogen fertilizers. The study also suggestsusing automatic differential harvesting as another promising actuation process for promoting the harvesting process over the field. Adoption of differential harvesting process is reported in Sethuramasamyraja [39], where infrared sensors were used over vineyards to analyze the quality of graph based on anthocyanin present in berries. The implementation is carried out as follows viz. i) anthocyanin contents of the grapes are sensed, ii) a certain level of the threshold for this content is considered to generate a quality map for this data, and iii) forwarding the generated map to the user (harvester).

- *Geomapping and Remote Sensing*: There are various forms of Geomapping and remote sensing approaches used towards PA (Kim et al.[40]). This approach leads to the generation of agroecological zones where different attributes are subjected to analysis (Muthoni et al. [41]). The imageries obtained from satellite images are studied for boundary delineation using feature extraction and image segmentation method (North et al. [42]). The existing study has also witnessed increased adoption of Sentinel-2 data in PA (Sharifi [43]) for analyzing nitrogen usage. Nitrogen is the essential input for PA has also been studied by Yao et al. [44] using an active crop sensor. Apart from these conventional approaches, the advanced integrated approach of drone technology and Internet-of-Things are also deployed in precision farming (Uddin et al. [45]). Another interesting study carried out by Xu et al. [46] has used data from cameras and terrestrial laser scanning to monitor crop health in PA. The majority of the approaches associated with Geomapping and remote sensing are associated with capturing the field image followed by performing analysis. Proximal sensing is most recently integrated with remote sensing from multiple sources to study the leaf area index (Asad et al. [47]). This work connects the health statistics of the leaf with the topographical map of the earth. This model has three distinct modules viz. i) data processing with semantic segmentation of ground images, ii) training using deep learning model, and iii) performing prediction. The study outcome suggests that it is capable of performing better prediction even with images with low resolution.

The current study has also discussed spectral feature usage, where the prime challenge is to address the issues associated with data collection and training. This issue is addressed in Ashourloo et al. [48], which carried out a comparative study of different variants of spectral bands. The outcome shows support vector machine to be useful for large scale of data using time-series approach. However, such an approach is less utilized for computing as well as predicting yield. This problem is addressed in the work of Fieuzal et al. [49] considering leaf area index. The data considered for this analysis is from synthetic aperture radar, where multiple sources are considered for analysis for evaluating a crop's dry mass. A similar study is also carried out by Zalite et al. [50], where time series is considered. The study limits its evaluation from the wetlands, which is another research challenge found in current times. The prime cause of this challenge is spectral similarity and the degree of heterogeneity involved in landmasses. A study to address this challenge is seen in Hempattarasuwan et al. [51], where quantitative analysis is carried out over historical data. The study implements a classification approach by combining three standard approaches, i.e., Mahalanobis distance, maximum likelihood, and decision tree. The outcome shows a decision tree to offer better classification performance. A study concerning leaf area index is also carried out by Pan et al. [52], where water content information is also used for modeling. The emphasis on water attributes was also seen in the study of Patil et al. [53]. The current study also claims that useful classification can be carried out using a PA's deep learning approach (Sun et al. [54]). From an approach perspective, the random forest has also registered itself to be assisting in the classification of satellite images of land (Zafari et al. [55]). In such an approach, a unique classifier is designed for constructing a similarity kernel. There are also studies where correlated factors, e.g., development stage and fractal dimension,are studied (Shen et al. [56]). Such study mainly explores different factors that affect production, i.e., soil background and different farming practices. A unique study carried out by Dong et al. [57] has used chlorophyll index for assessing the internal processing of crops in PA. The study carried out over simulated environment shows the potential linear correlation among different variants of vegetation index. The study contributes towards the impact of red edge reflectance associated with chlorophyll during photosynthesis. Such models emphasize the internal processing of plant nutrients but do not focus on balancing them. Balancing the nutrient demand is essential when it comes tothe management of agricultural land in PA. Such an approach was discussed by Gimenez et al. [58],where remotely sensed data is integrated with the model for land management. The study contributes towards yielding useful information associated with farm practices and balancing the nutrients demands on it. Existing studies have also evolved with a unique clustering approach on its features over the standard scale to assess the monitoring of crops in PA (Yuzugullu et al. [59]). The work carried out by Ali et al. [60] has developed a model for remote sensing where multitemporal attributes have been used for evaluating biomass. The study has used an integrated machine learning approach where neuro-fuzzy logic, neural network, and linear regression have been used over remotely sensed data to extract biomass estimates.

- *Automated Steering System:* The research work towards this approach is mainly associated with developing agricultural machinery to give them a direction towards its orientation. The existing system has used fuzzy logic (Duan et al. [61]), manual priority (Fu et al. [62]), renewable energy (Ghobadpour et al. [63]), proportional integral derivative (Liu et al. [64], Yin et al. [65]), designing electro-hydraulic circuit (Mungwongsa et al. [66]), field robots (Gonzalez-de-Santos et al. [67]), and automatic pilot system (Wang et al. [68]). The idea of the majority of such implementation orients about developing a system that can assist the agricultural machinery to accomplish specific objectives while farming. It reduced iterative human efforts andcan undertakea specific task that is not feasible for humans to carry out for a given constraint of extensive agricultural lands. However, most of the approaches are associated with hardware-based development, and less advancement is done on the computational model.

Table I highlights the summary of the most significant conventional PA-based approaches studied above-concerning issues, methodology, advantages, and limitations connected to them.

TABLE I. SUMMARIZATION OF THE EXISTING STUDIES IN PA [SPS: SATELLITE POSITIONING SYSTEM, VRT: VARIABLE RATE TECHNOLOGY, GRS: GEOMAPPING AND REMOTE SENSING, ASS: AUTOMATED STEERING SYSTEM

	Author	Problem	Methodology	Advantages	Limitation
SPS	Levoir et al. [25]	High complexity localization, occlusion of GPS	Autonomous GPS-based rover vehicle, image processing, statistics	Higher accuracy	Lacks standard benchmarking
	Rodriguez et al. [26]).	Data acquisition	Prototyping by integrating sensor and GPS	Assists in differential data acquisition	Lacks comparison with the existing system, does not consider signal unavailability in GPS
	Si et al. [27]	Calculating steering angle of farming vehicle	Prototyping with gyroscope, unscented Kalman filter, GPS	Higher accuracy	Involves higher computation to compute steering angle
	Dabove et al. [28]	Receiver effectiveness with GPS	Discussion of different variants of GPS-based receiver and antenna	Simplified discussion	It does not conclude the best performing receiving in adverse environmental condition
	Marucci et al. [29]	Effectiveness of using GNSS	An experimental model combining RTK with GNSS	Improved accuracy of trajectories	It does not deal with heterogeneous environments of farming
VRT	Nordblom et al. [33]	Search for the reason for waterlogging	Simulation-based study	Simplified probability model, risk analysis	Region-specific study
	Steffani [34]	Risk analysis of cotton production	Statistical modeling	Simplified risk analysis	Region-specific study
	Kweon et al. [35]	Testing of organic matter of soil	Prototyping, field study, sensors, linear regression (multivariate)	Comprehensive analysis	Computational complexity is higher and not addressed
	Colaco& Molin [36]	Fertilization of citrus	Discussion of variable rate fertilization, yield map	Reduction in input,	Study-specific to region and crop
	Nawar et al. [37]	Zone delineation management	Discussion of various techniques and their contribution	Pin-pointed findings to prove increased yield	It does not discuss the inclusion of high-end analytics
	Thomasson et al. [38]	Automation technologies	Discussion of robotics and automation in PA	Discusses the importance of robotics in PA	It does not discuss the significant approach
GRS	North et al. [42]	Boundary delineation	Image segmentation, feature extraction	Higher suitability towards the classification of land	Area-specific study
	Uddin et al. [45]	Health monitoring of crop	Drone with IoT, dynamic clustering of data	Wide applicability, cost-effective	Hypothetical model
	Xu et al. [46]	Health monitoring of crop	Scanning with terrestrial laser, cloud data	Higher precision	It does not support heterogeneous modeling
	Asad et al. [47]	Index area mapping of leaf	Deep learning	The prediction does not demand high image resolution	Iterative mechanism,
	Ashourloo et al. [48]	Data collecting during remote sensing	Time-series, support vector machine	Assists in involuntary crop mapping	Training time is higher.
	Fieuzal et al. [49]	Lack of well-sampled data in time series, analysis of leaf area index	Combined analysis of satellite data and agrometeorological data	Effective simulation of temporal feature	Study restricted to specific crop (sunflower)
	Hempattarasuwan et al. [51]	Wetland classification	Integrated classification approach	Decision tree found to offer higher accuracy	This leads to computational complexity
	Pan et al. [52]	Analysis of multispectral data	Integrating leaf area index and water content, neural network	Good accuracy	It does not include the environmental uncertainty factor
	Patil et al. [53]	Water productivity assessment	Energy balance for surface	Lower predictive errors	Specific to desert farming
	Zafari et al. [55]	Classification of land	Randomized tree, kernel	Able to solve high-dimensional data	Study-specific to support vector machine
	(Shen et al. [56]).	Crop type classification	Deep learning	Reliable map generation	Does not address the computational complexity of training.

	Dong et al. [57]	Assessing vegetation index	Algorithm for extracting reflectance of active chlorophyll	Capable of assessing the impact of vegetation impact	Study-specific to chlorophyll
	Gimenez et al. [58]	Classification of land usage	Integrating remotely sensed data with a model of land management	Increasing accuracy in the information of land usage	Increased processing time
	Ali et al. [60]	Biomass estimation	Machine learning	Enhanced estimation approach	Accuracy depends upon the amount of trained data, presence of anomalies from the satellite signal
ASS	Duan et al. [61]	Real-time control on machinery	Fuzzy Logic	Improve accuracy in steering	It depends upon ruleset construction
	Fu et al. [62], Liu et al. [64], Mungwongsa et al. [66]	Automated steering	Electro-hydraulic steering, sensor	Reduced response time	Lacks smart feature
	Ghobadpour et al. [63]	Automated steering	Renewable energy system	Discusses increasing trend	It does not highlight the effective approach
	Gonzalez-de-Santos et al. [67]	Intelligent farming	Robotics	Discusses autonomous robots	Research gap not identified
	Wang et al. [68], Yin et al. [65]	Autonomous robots	Embedded system	Good accuracy	No benchmarking

III. WSN IN PRECISION AGRICULTURE

With the advent of the dominant adoption of sensors, current research work towards PA has been revolutionized more toward incorporating smart sensing. One of the prime motivations towards this research trend is the increasing awareness of Internet-of-Things (IoT), where sensors are integral. IoT is one dominant research topic for improving agricultural yields (Kour and Arora [69]). It has contributed towards opening avenues for smart farming and PA, although there is some dominant research gap (Kour and Arora [69]). In this aspect, various forms of sensors have also been investigated towards PA, where it is found that support vector machine and random forest are dominant classification approaches (Kamath et al. [70]). Apart from this, there is also dedicated research work being carried out where machine learning approaches are claimed to optimize IoT performance in PA to facilitate predictive operation for farming.

With the adoption of various sensors for capturing field information, the data are forwarded using various IEEE standards of the family (e.g., 802.15.4/11 as seen in the work of Kone et al. [71]), which further forwards it to the gateway node and then to cloud where the application of analytics resides (Ahmed et al. [72]). The study offers some specific information that was not found in conventional PA-based approaches, e.g., i) energy being one of the practical constraintsof using sensors in PA, and ii) routing and topology is another essential operation, which is also challenged in adverse environmental condition. There are various MAC protocols in wireless sensor networks [72], but they do not combine to ensure downlink scheduling, multi-hop decisions, heterogeneous duty cycles, and traffic adaptive. To perform a full scenario to capture environment information of farming process, all this characteristic is demanded in IoT. The adoption of IoT technology in PA is depicted in Fig. 6. The figure shows how sensor devices, gateways, and WiFi technology integrated with cloud infrastructure enableIoT-PA ecosystems. There are basically severalwireless sensor nodes deployed in the farm and agriculture fields in rural regions. The sensor nodes capture significant events related to agriculture and send them to the cloud computing system via WiFi and gateway-based networking systems. The sensed data collected to the cloud is further stored and processed by an analytics engine and fog networking to enable framers managing farms to boost the quality and quantity of products and optimizes the cost associated with human labor required. However, in this scenario, the biggest impediment is a trade-off concerning supportability and efficiency between the protocols in IoT and Wireless Sensor Network (WSN).

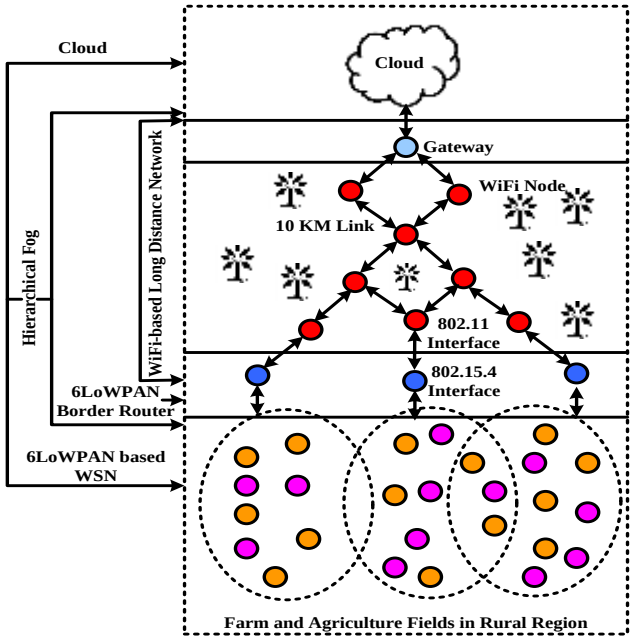


Fig. 6. Adoption of IoT in PA.

The most recent study carried out by Gulec et al. [73] has discussed improving the lifetime of WSN focusing on PA in a distributed environment. The study uses connected dominating sets as a backbone of communication in WSN considering harvester and regular sensors in farming. The study outcome is obtained from both experimental and simulated versions stating that the proposed system is energy efficient. Existing research

shows certain dedicated attempts to model WSN in PA with uniform sensory node distribution over the farming area. The work carried out by Bacco et al. [74] has developed a channel model that is used by the ground sensors to perform data transfer. However, the emphasis was more on the usage of IEEE standards and less on WSN. Adopting the heterogeneous sensor network is witnessed in Sylvian et al. [75] and Kaiwartya et al. [76]. In this work, prototyping is carried out using different sensors to capture different information associated with farm fields and crops.

Further, WSN also claims to offer a decision support system for facilitating water usage (Khan et al. [77]). A prototype is designed where the temperature is used for environmental monitoring in PA. The study analyzes the consumption of current while functioning over different degrees of temperatures. Importance over plant water is another investigation in the existing system, an essential part of the leaf sensing system in WSN with PA. The current study claims that the adoption of backscatter-based sensor nodes could enhance the PA performance from the perspective of power-saving (Daskalakis et al. [78]). The study has also used Morse code, which is computationally cost-effective for carrier signal modulation. Focus on power saving can also be implemented using non-orthonormal multiple access in WSN (Hu et al. [79]). The study outcome shows that this mechanism significantly controls outage probability and the rate of summed data.

The majority of the existing studies emphasize estimating soil parameters in PA; however, the modeling attributes are less emphasized towards power. A study on such issues is carried out by Estrada-Lopez et al. [80], where a WSN topology is constructed using both cloud and IoT considering soil parameters. The data analysis is carried out by an artificial neural network followed by using a unique power management scheme. Apart from the terrestrial application, the adoption of WSN is also carried out over the underground ecosystem. Salam et al. [81] have developed such a system to model channel impulse. The study has also analyzed various time-domain attributes, e.g., gain in multipath power, channel capacity, delay, etc. A study on a similar direction towards the underground ecosystem is also investigated by Castellanos et al. [82], where soil parameters are collected using a narrow-band communication scheme of Long Term Evolution (LTE). The study uses unmanned aerial vehicles to collect data from underground sensors over the potato crop field. Another study of the underground ecosystem is carried out by Sambo et al. [83], where a path loss model is presented along with predictive framework development using a complex dielectric constant.

Deployment of WSN in PA was also claimed to enhance productivity by using dataloggers and actuators (Lozoya et al. [84]). The current study of WSN is also focused on incorporating intelligence in the process of irrigation in PA. The work carried out by Jamroen et al. [85] has developed an irrigation scheduling mechanism using fuzzy logic in WSN. The outcome witnessed an increase in crop yield. The current study also discusses the usage of WSN for assisting in localizing in PA. Sahota and Kumar [86] have implemented a model where the received signal strength has been used for distribution over WSN. The study develops a node localization model considering distance propagation invarious degrading effective over the signal considering fading and path loss model. The study contributes to predicting loss in nitrogen. A similar received signal strength-based approach for assisting in localization has also been carried out by Abouzar et al. [87]. This study has used a spanning tree for developing belief propagation.

A unique concept towards promoting energy harvesting in PA is discussed by Konstantopoulos et al. [88]. According to this study, the electric potential produced within a plant is used as a power source for WSN. The study uses nonnegative matrix factorization to process this electric potential signal. From the viewpoint of power saving, it is also found that data registers' frequency plays a crucial role. The energy-saving in WSN can be facilitated using this data register frequency variation to impact PA (Santos and Cugnasca [89]). Another essential factor to be considered is the presence of partitioned sensors in PA, which leads to disruption in the network. The work carried out by Maheswararajah et al. [90] hypothesizes that the presence of such nodes leads to noise in the measurement. A Kalman-filter-based optimization strategy is developed to restore such nodes. Existing literature further hypothesizes that monitoring environmental values is essential when deploying WSN in PA. The work of Kampianakis et al. [91] has presented a prototype that employs the networking principle of sensor nodes (especially modulation of analog frequency) along with software-defined radio.

The summary of the practical approaches in WSN in PA is tabulated in Table II.

TABLE II. SUMMARY OF WSN-BASED PA APPROACHES

Author	Problem	Methodology	Advantages	Limitation
Gulec et al. [73]	Network lifetime	Connected dominating sets, solar energy harvesting	Reduced energy consumption	Lacks considering different resource retention
Bacco et al. [74]	Coverage and Connectivity	Channel model	Simplified design	Only focus on IEEE 802.15.4 usage
Sylvian et al. [75], Kaiwartya et al. [76]	Health monitoring of crops	Multi-sensor prototype	Effective field measurement	Lacks benchmarking
Khan et al. [77]	Water utilization in the farming area	Decision support system	Higher accuracy	does not consider energy factor
Daskalakis et al. [78]	Plant water monitoring	Backscatter	Power saving	Cost is still incurred in the usage of multiple equipments
Hu et al. [79]	Enhancing Network lifetime	Non-Orthonormal Multiple Access	Reduces outage probability	Not applicable for the sparse network.
Estrada-Lopez et al. [80]	Power management, soil parameter estimation	Artificial neural network, cloud, IoT	Enhanced reliability and better system performance	The study uses a specific sensor node, which demands more training for accuracy.
Salam et al. [81]	Underground channel development in WSN for assessing soil health	Assessing impulse response	Approach with practical constraints, reduced energy depletion, reduced delay	The routing aspect is not considered in WSN
Castellanos et al. [82]	Computation of link quality	Narrow-band communication, path loss model	Applicable for both under and above ground operation	It does not ensure scalability owing to the defined range.
Sambo et al. [83]	Underground monitoring in PA	Path loss model, predictive	Higher accuracy	Performs a highly iterative operation
Jamroen et al. [85]	Irrigation scheduling	Fuzzy logic	Reduces energy consumption, increased crop yield	Increased dynamic attributes may cause an increase in fuzzy rules
Sahota and Kumar [86], Abouzar et al. [87]	A crop network architecture in PA	Received signal strength, maximum likelihood	Resistive against multipath fading	Cannot sustain over intermittent links in WSN
Konstantopoulos et al. [88]	Energy harvesting	Nonnegative matrix factorization	Highly cost useful energy source	Workability over extensive, dense, and uncertain network is not evaluated
Maheswararajah et al. [90]	The partitioned node in WSN	Kalman Filter	Reduced error rate	Error computation is resource-dependent and hence not scalable for large networks.
Kampianakis et al. [91]	Environmental monitoring	Prototyping, software-defined radio	Higher precision	It demands excessive power consumption

IV. REVIEWING RESEARCH TREND

From the perspective of the global trend, it is seen that IoT, along with the inclusion of software and different variants of sensing technology, is going to minimize the skilled labors in agriculture in the coming days. The global market is not consistently evolving with the rise of real-time kinetic technology, remote sensing technology, networking, variable rate technology, robotics, and fertilizers and sprayer controllers.

A. Trend in PA Research

The last decade has witnessed approximately 1710 research papers in PA approaches while only 230 are found to be journals in IEEE Xplore digital library. A nearly similar trend is found in other reputed publishers like ACM digital library, Springer, ScienceDirect, and Elsevier. There are very few studies towards automated steering systems, while more studies are populated in the adoption of satellite positioning systems (GPS, GNSS). Not much work is carried out towards variable rate technology. However, some potential work in a large number has been carried out towards remote sensing and soil mapping. More inclination is seen towards remote sensing approaches using hyperspectral images or other equivalent forms of images from an unmanned flying object (drones). However, the trend is more on adopting a single crop field is extensively more investigated, and multi-crop land is less found in consideration, which could impedeupcoming research work. Agriculture 4.0 is an upcoming standard for automating PA; however, studies show few implementations associated with such upcoming standard formulation. Image processing remains the dominant approach, and its adoption is consistently increasing; however, there is a shift of this approach with data-centric technologies in IoT.

B. Trend in Technological Adoption

The present scenario of implementation in PA is highly scattered. More work is carried out using prototyping, and less mathematical or computational modeling is noticed. Adoption of machine learning or artificial intelligence is also found to be less prominent in this aspect. Although machine learning has been used in existing studies, it is not evaluated fromits

computational complexity. The engineering area, e.g., robotics, embedded system, machinery compilation, etc. is more focused, limiting investigation strength and giving less exposure to unknown challenges in PA. Adopting IoT and WSN has just started its research work, and it has more way to go to achieve its state of maturity as a research standard model. The development of a test-bed for analyzing farming data is another inclusive research trend in PA.

C. Trend in Target Issues

The issues mainly considered in the existing system in PA are mainly associated with environmental monitoring. The existing research trend is also to consider a specific issue connected with a specific crop, making the model heavily case-specific and less applicable to different environments. Data acquisition is another target issue considered in the existing research trend in PA. Different techniques have been carried out towards acquiring data. However, less emphasis is offered to analyze this collected data. The trend towards analytics over multi-crop land is less found. Adopting sensors integrated with different networking principles also assists in data acquisition; however, there are various open-end challenges associated, e.g., non-inclusion of the energy model makes such a solution limited to theoretical concepts.

V. Discussion and Perspective

Based on the observation being carried out towards conventional approaches used in PA and the upcoming adoption of WSN in PA, it is noticed that there are various concluding remarks associated with the overall techniques used in PA. This section briefs about the learning outcomes of the proposed review work as follows:

- *A tradeoff between Demands and Available Technology*: A closer look into the available approaches shows that PA needs to consider multiple attributes simultaneously, e.g., soil health, plant-related features, surrounding environment, and weather. There are many more sub-attributes for this core attribute, which require equal attention for improved crop cultivation and environmental risk reduction. All the existing approaches using a conventional approach or WSN based approaches use only a limited number of such attributes in modeling its PA. On the other side, there has been an immense advancement in prototyping as well as computational modeling. However, prototyping is the most dominant approach in PA in existing studies. Hence, the demand to offer productive PA performance is immensely more which are not found to be considered while modeling with existing technological advancement.
- *Lack of Uncertainty-based Modelling*: There are various attributes like crop health, rainfall, temperature, soil health, etc. they are stochastic. Existing approaches focus on modeling predefined ecosystems, which is more or less impractical than real-world scenarios. There isa various uncertain scenario that could develop either using conventional or WSN based approaches, e.g. rate of energy depletion,incoming streamed data, mobile of machinery, occlusion in GPS-based data, etc. Until and unless such uncertainty conditions are not included in the modeling, the outcome may eventually result in outliers. Apart from this, various studies where machine learning has been used do not consider this, leading to its solution inapplicable to real-time application.
- *Use Case Specific Study*: Almost all the existing PA approaches have considered a specific use case of crop or study environment (e.g., soil health, water, temperature, etc.). On the other side, the conventional study approach has focused on the adoption of specific machinery. The modeling is carried out considering a specific form of crops using any of the approaches in PA. This means that there is no generalized algorithm to solve a similar problem when environmental variables change. It also incurs more cost when it comes to deploying commercial products and their adoption. It is only cost-effective of a simplified model (or product) that can address multiple PA problems altogether.
- *Less Emphasis over Routing*: Routing or deploying a communication protocol is significant usingthe larger farming area with challenging communication scenarios (e.g., forest, terrain, etc.). It is already observed that the adoption of the hybrid approach is the most effective one to mitigate the limitation of single-approach. For example, GPS integrated with sensor nodes or drones could offer more effective data capture than considering any of them. This also means that there isa good possibility of hybridizing different types of machinery and different nodes to facilitate an effective data transmission in PA. However, this challenge can be addressed if a unique routing protocol is designed and developed for such a scenario. No studies are being carried out in evolving a novel routing scheme in PA;instead, it reuses the adopted techniques' routing scheme. This also offers more impediments towards data transmission when the farm environment is subjected to priority-based data transmission or exercising specific time-critical applications.
- *IoT and WSN still in the Nascent Stage*: IoT is slowly making its entry from the roof of research and development to the commercial world. Apart from this, the study shows that most PA approaches have a deployment of sensor nodes for soil mapping, remote sensing, etc. (conventional approach in PA), but they do not have a deployment of WSN, which makes a network of sensors. With the inclusion of automation standard 4.0, there is a need for smart farming using IoT, which is still under development. Apart from this, WSN is an integral part of IoT. However, there has been immense work towards addressing multiple problems in WSN in past decades, and their solutions are not directly applicable in IoT. There is always a tradeoff between IoT and WSN with the inclusion of IoT based routing scheme and WSN based routing scheme that requires smooth integration. Hence,

current approaches in WSN on PA aresignificantly less and insignificant in contrast to conventional PA approaches.

- *Lack of inclusion of Resource*: Sensor nodes of any form are characterized by the limited capability of processing as well as they have limited availability of resources too (e.g., memory, channel capacity, energy, etc.). None of the existing studies where WSN is considered in PA has any inclusion of novel resource management model exclusively focusing on constraints associated with PA's farming environment. Without the inclusion of the resource factor, modeling any solution will be more impractical.
- *Few Studies towards Optimization*: By optimization, it can represent a technique that offers increased performance yield with low inclusion/dependencies of resources. Machine learning has been used for this purpose to some extent. At present, many optimization-based approaches fit on solving various problems associated with PA. A closer look into the existing system also shows that it does not ensure computational cost-effectiveness in its algorithm. Hence, the adoption of appropriate optimization techniques is highly demanded.

VI. Conclusion

The manuscript discusses the PA approaches and techniques that are mainly associated with implementing a management scheme towards facilitating effective responses toward crops, measurement, and observation towards animals and fields. Adoption of PA leads to enhanced yields in the crop, cost reduction, and process input optimization. However, there are various challenges associated with it. There is an inclusion of higher initial capital to implement PA in real-time, and such investment is carried out for long-term plans. In order to reach the PA implementation maturity stage, several years may be consumed prior to even possessing adequate data to implement even the conventional approaches completely. The final challenge in PA implementation is its data aggregation followed by an analysis, which could be an extensively demanding task. Based on the presented findings of existing research work, it could be just said that effective implementation of PA demands i) precise management, ii) identification and adoption of appropriate technology, and iii) data.

1) Overall summary: The essential findings of the proposed study are summarized as follows: i) existing approaches of PA has an increasing concern over interoperability of different innovative systems and tools, ii) adoption of PA by ordinary farmers will be a big task as the technologies involved in it are highly advanced and require a thorough knowledge of it, iii) despite various studies using IoT, narrowband, GPS, WSN, etc., coverage and connectivity in rural areas will be a potentiallytricky task, iv) An appropriate PA implementation leads to generate a massive score of big farming data which is impossible to analyze from a single data point in the crop field. With the increasing adoption of multi-crop land, there will be massive growth of data and understanding the significance and priority of such data will be near to impossible for average farmers in existing times, v) IoT and WSN is the most promising technology in PA, but adoption of current schemes only induces scalability problems along with troublesome configuration issues, vi) there is a lack of mathematical modelling seen in the existing system using WSN, which has better future scope.

2) Future work: The future direction of work will consider adopting IoT and WSN, which is the most demanding upcoming technology for reshaping the existing system to Farming 4.0. In this context, the next work is to design and develop an IoT scenario with multi-crop land powered by heterogeneous WSN. The focus will be first to include all real-time constraints, e.g., energy, coverage and connectivity, resource management of the sensors. The secondary focus is to formulate a novel routing scheme thatoffers flexibility, scalability, and resource efficiency. It is also necessary to perform the complete modeling using the computational model, considering its applicability to practical world scenarios. The inclusion of multiple challenging test-bed and an effective validation technique could further offer more reliability to PA's upcoming solution.

References

[1] Vecchio, Yari, Marcello De Rosa, Felice Adinolfi, Luca Bartoli, and Margherita Masi. “Adoption of precision farming tools: A context-related analysis.” Land Use Policy 94 (2020): 104481.

[2] Abobatta, Waleed Fouad. “Precision Agriculture: A New Tool for Development.” In Precision Agriculture Technologies for Food Security and Sustainability, IGI Global (2021) pp. 23-45.

[3] Torky, Mohamed, and Aboul Ella Hassanein. “Integrating blockchain and the internet of things in precision agriculture: Analysis, opportunities, and challenges.” Computers and Electronics in Agriculture (2020): 105476.

[4] Li, Wenjing, Beth Clark, James A. Taylor, Helen Kendall, Glyn Jones, Zhenhong Li, Shan Jin, et al. “A hybrid modelling approach to understanding adoption of precision agriculture technologies in Chinese cropping systems.” Computers and Electronics in Agriculture 172 (2020): 105305.

[5] Priya, Rashmi, and Dharavath Ramesh. “ML based sustainable precision agriculture: A future generation perspective.” Sustainable Computing: Informatics and Systems 28 (2020): 100439.

[6] Klerkx, Laurens, and David Rose. “Dealing with the game-changing technologies of Agriculture 4.0: How do we manage diversity and responsibility in food system transition pathways?” Global Food Security 24 (2020): 100347.

[7] Vecchio, Yari, Marcello De Rosa, Felice Adinolfi, Luca Bartoli, and Margherita Masi. “Adoption of precision farming tools: A context-related analysis.” Land Use Policy 94 (2020): 104481.

[8] Symeonaki, Eleni, Konstantinos Arvanitis, and DimitriosPiromalis. “A context-aware middleware cloud approach for integrating precision farming facilities into the IoT toward agriculture 4.0.” Applied Sciences 10, no. 3 (2020): 813.

[9] Boursianis, Achilles D., Maria S. Papadopoulou, Panagiotis Diamantoulakis, Aglaia Liopa-Tsakalidi, PantelisBarouchas, George Salahas, George Karagiannidis, Shaohua Wan, and Sotirios K. Goudos. “Internet of things (IoT) and agricultural unmanned aerial vehicles (UAVs) in smart farming: a comprehensive review.” Internet of Things (2020): 100187.

[10] Singh, Ritesh Kumar, MichielAernouts, Mats De Meyer, Maarten Weyn, and Rafael Berkvens. “Leveraging LoRaWAN technology for precision agriculture in greenhouses.” Sensors 20, no. 7 (2020): 1827.

[11] Ofori, Martinson, and Omar El-Gayar. “Drivers and challenges of precision agriculture: a social media perspective.” Precision Agriculture (2020): 1-26.

[12] Cabezas-Cabezas, Roberto, Jomar Guzmán-Seraquive, Kevin Gómez-Gómez, and Corima Martínez-Villacís. "Integrated System for the Improvement of Precision Agriculture Based on IoT." In International Conference on Technologies and Innovation, pp. 123-136. Springer, Cham, 2020.

[13] Groher, Tanja, Katja Heitkämper, Achim Walter, Frank Liebisch, and Christina Umstätter. “Status quo of adoption of precision agriculture enabling technologies in Swiss plant production.” Precision Agriculture 21, no. 6 (2020): 1327-1350.

[14] Keswani, Bright, Ambarish G. Mohapatra, Poonam Keswani, Ashish Khanna, Deepak Gupta, and Joel Rodrigues. “Improving weather dependent zone specific irrigation control scheme in IoT and big data enabled self-driven precision agriculture mechanism.” Enterprise Information Systems 14, no. 9-10 (2020): 1494-1515.

[15] Demestichas, Konstantinos, and Emmanouil Daskalakis. “Data Lifecycle Management in Precision Agriculture Supported by Information and Communication Technology.” Agronomy 10, no. 11 (2020): 1648.

[16] Nel, Werner, Jan C. Boelhouwers, Carl-Johan Borg, Julian H. Cotrina, Christel D. Hansen, Natalie S. Haussmann, David W. Hedding et al. “Earth science research on Marion Island (1996–2020): a synthesis and new findings.” South African Geographical Journal (2020): 1-21.

[17] Wang, Nan, Weiming Cheng, Baixue Wang, Qiangyi Liu, and Chenghu Zhou. “Geomorphological regionalization theory system and division methodology of China.” Journal of Geographical Sciences 30, no. 2 (2020): 212-232.

[18] Groher, Tanja, Katja Heitkämper, Achim Walter, Frank Liebisch, and Christina Umstätter. “Status quo of adoption of precision agriculture enabling technologies in Swiss plant production.” Precision Agriculture 21, no. 6 (2020): 1327-1350.

[19] Radoglou-Grammatikis, Panagiotis, Panagiotis Sarigiannidis, Thomas Lagkas, and IoannisMoscholios. “A compilation of UAV applications for precision agriculture.” Computer Networks 172 (2020): 107148.

[20] Bucci, Giorgia, Deborah Bentivoglio, Matteo Belletti, and Adele Finco. “Measuring the farm’s profitability after the adoption of Precision Agriculture Technologies: A case study research from Italy.” Acta IMEKO 9, no. 3.

[21] Raj, Rahul, Soumyashree Kar, Rohit Nandan, and AdinarayanaJagarlapudi. “Precision agriculture and unmanned aerial vehicles (UAVs).” In Unmanned Aerial Vehicle: Applications in Agriculture and Environment, pp. 7-23. Springer, Cham, 2020.

[22] Balasundram, Siva K., Kamlesh Golhani, Redmond R. Shamshiri, and Ganesan Vadamalai. “Precision agriculture technologies for management of plant diseases.” In Plant Disease Management Strategies for Sustainable Agriculture through Traditional and Modern Approaches, pp. 259-278. Springer, Cham, 2020.

[23] Pang, Guowei, Qinke Yang, Chunmei Wang, Rui Li, and Lu Zhang. “Quantitative assessment of the influence of terrace and check dam construction on watershed topography.” Frontiers of Earth Science 14 (2020): 360-375.

[24] Turcian, Daniel, ValerDolga, Darius Turcian, and Cristian Moldovan. “Fusion Sensors Experiment for Active Cruise Control.” In Joint International Conference of the International Conference on Mechanisms and Mechanical Transmissions and the International Conference on Robotics, pp. 432-443. Springer, Cham, 2020.

[25] LeVoir, Samuel J., Peter A. Farley, Tao Sun, and Chong Xu. “High-Accuracy Adaptive Low-Cost Location Sensing Subsystems for Autonomous Rover in Precision Agriculture.” IEEE Open Journal of Industry Applications 1 (2020): 74-94.

[26] Rodriguez, Roberto, Daniel M. Jenkins, and James JK Leary. “Design and validation of a GPS logger system for recording aerially deployed herbicide ballistic technology operations.” IEEE Sensors Journal 15, no. 4 (2014): 2078-2086.

[27] Si, Jiaqian, YanxiongNiu, Jiazhen Lu, and Hao Zhang. “High-Precision Estimation of Steering Angle of Agricultural Tractors Using GPS and Low-Accuracy MEMS.” IEEE Transactions on Vehicular Technology 68, no. 12 (2019): 11738-11745.

[28] Dabove, P., and A. M. Manzino. “GPS mass-market receivers for precise farming.” In Proceedings of IEEE/ION PLANS 2014, pp. 472-477. 2014.

[29] Marucci, Alvaro, Andrea Colantoni, Ilaria Zambon, and Gianluca Egidi. "Precision farming in hilly areas: The use of network RTK in GNSS technology." Agriculture 7, no. 7 (2017): 60.

[30] Stombaugh, Timothy. “Satellite-based Positioning Systems for Precision Agriculture.” Precision agriculture basics (2018): 25-35.

[31] Saiz-Rubio, Verónica, and Francisco Rovira-Más. “From smart farming towards agriculture 5.0: A review on crop data management.” agronomy 10, no. 2 (2020): 207.

[32] Ayaz, Muhammad, Mohammad Ammad-Uddin, Zubair Sharif, Ali Mansour, and El-Hadi M. Aggoune. “Internet-of-Things (IoT)-based smart agriculture: Toward making the fields talk.” IEEE Access 7 (2019): 129551-129583.

[33] Nordblom, Thomas L., Timothy R. Hutchings, Sosheel S. Godfrey, and Cassandra R. Schefe. “Precision variable rate nitrogen for dryland farming on waterlogging Riverine Plains of Southeast Australia?.” Agricultural Systems 186 (2020): 102962.

[34] Stefanini, Melissa Reynolds. “Effects of optical sensing and variable rate technology on nitrogen fertilizer use, lint yields, and profitability in cotton production.” (2015).

[35] Kweon, G.; Lund, E.; Maxton, C. Soil organic matter and cation-exchange capacity sensing with on-the-goelectrical conductivity and optical sensors. Geoderma 2013, 199, 80–89

[36] Colaço, A.F.; Molin, J.P. Variable rate fertilization in citrus: A long term study. Precis. Agric. 2017, 18, 169–191.

[37] Nawar, S.; Corstanje, R.; Halcro, G.; Mulla, D.; Mouazen, A.M. Delineation of Soil Management Zones for Variable-Rate Fertilization. Adv. Agron. 2017, 143, 175–245

[38] Thomasson, A.; Baillie, C.; Antile, D.; Lobsey, C.; McCarthy, C. Autonomous Technologies in Agricultural Equipment: A Review of the State of the Art. In Proceedings of the 2019 Agricultural Equipment Technology Conference, Louisville, KY, USA, 11–13 February 2019; American Society of Agricultural and Biological Engineers: St. Joseph, MI, USA, 2019. ASABE Publication Number 913C0119.

[39] Sethuramasamyraja, B. Precision Ag Research at California State University, Fresno. Resource 2017, 24, 18–19.

[40] Kim, Jeongeun, Seungwon Kim, Chanyoung Ju, and Hyoung Il Son. “Unmanned aerial vehicles in agriculture: A review of perspective of platform, control, and applications.” IEEE Access 7 (2019): 105100-105115.

[41] Muthoni, Francis. “Spatial-Temporal Trends of Rainfall, Maximum and Minimum Temperatures Over West Africa.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020): 2960-2973.

[42] North, Heather C., David Pairman, and Stella E. Belliss. “Boundary delineation of agricultural fields in multitemporal satellite imagery.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, no. 1 (2018): 237-251.

[43] Sharifi, Alireza. “Using sentinel-2 data to predict nitrogen uptake in maize crop.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020): 2656-2662.

[44] Yao, Yinkun, Yuxin Miao, Qiang Cao, Hongye Wang, Martin L. Gnyp, Georg Bareth, Rajiv Khosla, Wen Yang, Fengyan Liu, and Cheng Liu. “In-season estimation of rice nitrogen status with an active crop canopy sensor.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7, no. 11 (2014): 4403-4413.

[45] Uddin, Mohammad Ammad, Muhammad Ayaz, El-Hadi M. Aggoune, Ali Mansour, and Denis Le Jeune. “Affordable broad agile farming system for rural and remote area.” IEEE Access 7 (2019): 127098-127116.

[46] Xu, Lijun, Teng Xu, and Xiaolu Li. “Corn Seedling Monitoring Using 3-D Point Cloud Data From Terrestrial Laser Scanning and Registered Camera Data.” IEEE Geoscience and Remote Sensing Letters 17, no. 1 (2019): 137-141.

[47] Asad, Muhammad Hamza, and Abdul Bais. "Crop and Weed Leaf Area Index Mapping Using Multi-Source Remote and Proximal Sensing." IEEE Access 8 (2020): 138179-138190.

[48] Ashourloo, Davoud, Hamid Salehi Shahrabi, Mohsen Azadbakht, Hossein Aghighi, Ali Akbar Matkan, and Soheil Radiom. "A novel automatic method for alfalfa mapping using time series of landsat-8 OLI Data." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11, no. 11 (2018): 4478-4487.

[49] Fieuzal, Remy, Claire Marais Sicre, and Frederic Baup. "Estimation of sunflower yield using a simplified agrometeorological model controlled by optical and SAR satellite data." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, no. 12 (2017): 5412-5422.

[50] Zalite, Karlis, Oleg Antropov, JaanPraks, Kaupo Voormansik, and Mart Noorma. "Monitoring of agricultural grasslands with time series of X-band repeat-pass interferometric SAR." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9, no. 8 (2015): 3687-3697.

[51] Hempattarasuwan, Nuttiga, George Christakos, and Jiaping Wu. "Changes of Wiang Nong Lom and Nong Luang Wetlands in Chiang Saen Valley (Chiang Rai Province, Thailand) During the Period 1988–2017." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, no. 11 (2019): 4224-4238.

[52] Pan, Haizhu, Zhongxin Chen, Jianqiang Ren, He Li, and Shangrong Wu. "Modeling winter wheat leaf area index and canopy water content with three different approaches using Sentinel-2 multispectral instrument data." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, no. 2 (2018): 482-492.

[53] Patil, Virupakshagowda C., Khalid A. Al-Gaadi, RangaswamyMadugundu, ElKamil HM Tola, SamyMarey, Ali Aldosari, Chandrashekhar M. Biradar, and Prasanna H. Gowda. "Assessing agricultural water productivity in desert farming system of Saudi Arabia." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8, no. 1 (2014): 284-297.

[54] Sun, Ziheng, Liping Di, Hui Fang, and Annie Burgess. "Deep Learning Classification for Crop Types in North Dakota." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020): 2200-2213.

[55] Zafari, Azar, Raul Zurita-Milla, and Emma Izquierdo-Verdiguier. "Land cover classification using extremely randomized trees: A kernel perspective." IEEE geoscience and remote sensing letters 17, no. 10 (2019): 1702-1706.

[56] Shen, Yonglin, Liping Di, Genong Yu, and Lixin Wu. "Correlation between corn progress stages and fractal dimension from MODIS-NDVI time series." IEEE Geoscience and Remote Sensing Letters 10, no. 5 (2013): 1065-1069.

[57] Dong, Taifeng, Jihua Meng, Jiali Shang, Jiangui Liu, and Bingfang Wu. "Evaluation of chlorophyll-related vegetation indices using simulated Sentinel-2 data for estimation of crop fraction of absorbed photosynthetically active radiation." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8, no. 8 (2015): 4049-4059.

[58] Giménez, Marta Gómez, Raniero Della Peruta, Rogier de Jong, Armin Keller, and Michael E. Schaepman. "Spatial differentiation of arable land and permanent grassland to improve a land management model for nutrient balancing." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9, no. 12 (2016): 5655-5665.

[59] Yuzugullu, Onur, EsraErten, and Irena Hajnsek. "Rice growth monitoring by means of X-band co-polar SAR: Feature clustering and BBCH scale." IEEE Geoscience and Remote Sensing Letters 12, no. 6 (2015): 1218-1222.

[60] Ali, Iftikhar, Fiona Cawkwell, Edward Dwyer, and Stuart Green. "Modeling managed grassland biomass estimation by using multitemporal remote sensing data—A machine learning approach." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, no. 7 (2016): 3254-3264.

[61] Duan, Jie, Lei Zhang, Zhaoiinz Zhang, Jinzbo Zhao, and Yan Jiang. "Research on Automatic Steering System of Agricultural Machinery Based on Fuzzy Neural Network." In 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pp. 1602-1605. IEEE, 2018.

[62] Fu, Weiqiang, Guangwei Wu, Yue Cong, You Li, and Zhijun Meng. "Development of tractor automatic steering system with manual priority function." In 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 555-559. IEEE, 2015.

[63] Ghobadpour, Amin, LoïcBoulon, Hossein Mousazadeh, Ahmad Sharifi Malvajerdi, and Shahin Rafiee. "State of the art of autonomous agricultural off-road vehicles driven by renewable energy systems." Energy Procedia 162 (2019): 4-13.

[64] Liu, Jin-yi, Jing-quan Tan, En-rong Mao, Zheng-he Song, and Zhong-xiang Zhu. "Proportional directional valve based automatic steering system for tractors." Frontiers of Information Technology & Electronic Engineering 17, no. 5 (2016): 458-464.

[65] Yin, Chengqiang, Shourui Wang, Jie Gao, Ling Zhao, and Hequan Miao. "Steering tracking control based on assisted motor for agricultural tractors." International Journal of Control, Automation and Systems 17, no. 10 (2019): 2556-2564.

[66] Mungwongsa, Anon, KhwantriSaengprachatanarug, and Thana Radpukdee. "Design of an Automatic Steering System in a Small Farm Tractor." In 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC), pp. 224-229. IEEE, 2018.

[67] Gonzalez-de-Santos, Pablo, Roemi Fernández, Delia Sepúlveda, Eduardo Navas, Luis Emmi, and Manuel Armada. "Field Robots for Intelligent Farms—Inhering Features from Industry." Agronomy 10, no. 11 (2020): 1638.

[68] Wang, Maoli, Yongwei Tang, Huijuan Hao, Fengqi Hao, and Junfei Ma. "The design of agricultural machinery autonomous navigation system based on Linux-ARM." In 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pp. 1279-1282. IEEE, 2016.

[69] Kour, Vippon Preet, and Sakshi Arora. "Recent Developments of the Internet of Things in Agriculture: A Survey." IEEE Access 8 (2020): 129924-129957.

[70] Kamath, Radhika, Mamatha Balachandra, and Srikanth Prabhu. "Raspberry pi as visual sensor nodes in precision agriculture: A study." IEEE Access 7 (2019): 45110-45122.

[71] Kone, Cheick Tidjane, AbdelhakimHafid, and Mustapha Boushaba. "Performance management of IEEE 802.15. 4 wireless sensor network for precision agriculture." IEEE Sensors Journal 15, no. 10 (2015): 5734-5747.

[72] Ahmed, Nurzaman, Debashis De, and Iftekhar Hussain. "Internet of Things (IoT) for smart precision agriculture and farming in rural areas." IEEE Internet of Things Journal 5, no. 6 (2018): 4890-4899.

[73] Gulec, Omer, ElifHaytaoglu, and SezaiTokat. "A Novel Distributed CDS Algorithm for Extending Lifetime of WSNs With Solar Energy Harvester Nodes for Smart Agriculture Applications." IEEE Access 8 (2020): 58859-58873.

[74] Bacco, Manlio, Andrea Berton, Alberto Gotta, and Luca Caviglione. "IEEE 802.15. 4 air-ground UAV communications in smart farming scenarios." IEEE Communications Letters 22, no. 9 (2018): 1910-1913.

[75] Sylvain, Matthieu, Francis Lehoux, Steeve Morency, Félix Faucher, Eric Bharucha, Denise M. Tremblay, Frédéric Raymond et al. "The EcoChip: A wireless multi-sensor platform for comprehensive environmental monitoring." IEEE transactions on biomedical circuits and systems 12, no. 6 (2018): 1289-1300.

[76] Kaiwartya, Omprakash, Abdul Hanan Abdullah, Yue Cao, Ram Shringar Raw, Sushil Kumar, Daya Krishan Lobiyal, Ismail FauziIsnin, Xiulei Liu, and Rajiv Ratn Shah. "T-MQM: Testbed-based multi-metric quality measurement of sensor deployment for precision agriculture—A case study." IEEE Sensors Journal 16, no. 23 (2016): 8649-8664.

[77] Khan, Rahim, Ihsan Ali, Muhammad Zakarya, Mushtaq Ahmad, Muhammad Imran, and Muhammad Shoaib. "Technology-assisted decision support system for efficient water utilization: a real-time test-bed for irrigation using wireless sensor networks." IEEE Access 6 (2018): 25686-25697.

[78] Daskalakis, Spyridon Nektarios, George Goussetis, Stylianos D. Assimonis, Manos M. Tentzeris, and Apostolos Georgiadis. "A uW

backscatter-morse-leaf sensor for low-power agricultural wireless sensor networks." IEEE Sensors Journal 18, no. 19 (2018): 7889-7898.

[79] Hu, Zeng, Longqin Xu, Liang Cao, Shuangyin Liu, Zhijie Luo, Jing Wang, Xiangli Li, and Lu Wang. "Application of non-orthogonal multiple access in wireless sensor networks for smart agriculture." IEEE Access 7 (2019): 87582-87592.

[80] Estrada-López, Johan J., Alejandro A. Castillo-Atoche, Javier Vázquez-Castillo, and Edgar Sánchez-Sinencio. "Smart soil parameters estimation system using an autonomous wireless sensor network with dynamic power management strategy." IEEE Sensors Journal 18, no. 21 (2018): 8913-8923.

[81] Salam, Abdul, Mehmet C. Vuran, and Suat Irmak. "A Statistical Impulse Response Model Based on Empirical Characterization of Wireless Underground Channels." IEEE Transactions on Wireless Communications 19, no. 9 (2020): 5966-5981.

[82] Castellanos, German, Margot Deruyck, Luc Martens, and Wout Joseph. "System assessment of WUSN using NB-IoT UAV-aided networks in potato crops." IEEE Access 8 (2020): 56823-56836.

[83] Sambo, Damien Wohwe, Anna Forster, Blaise Omer Yenke, IdrissaSarr, Bamba Gueye, and Paul Dayang. "Wireless underground sensor networks path loss model for precision agriculture (WUSN-PLM)." IEEE Sensors Journal 20, no. 10 (2020): 5298-5313.

[84] Lozoya, Camilo, Alberto Aguilar, and Carlos Mendoza. "Service oriented design approach for a precision agriculture datalogger." IEEE Latin America Transactions 14, no. 4 (2016): 1683-1688.

[85] Jamroen, Chaowanan, PreechaKomkum, ChanonFongkerd, and WipaKrongpha. "An Intelligent Irrigation Scheduling System Using Low-Cost Wireless Sensor Network Toward Sustainable and Precision Agriculture." IEEE Access 8 (2020): 172756-172769.

[86] Sahota, Herman, and Ratnesh Kumar. "Maximum-likelihood sensor node localization using received signal strength in multimedia with multipath characteristics." IEEE Systems Journal 12, no. 1 (2016): 506-515.

[87] Abouzar, Pooyan, David G. Michelson, and Maziyar Hamdi. "RSSI-based distributed self-localization for wireless sensor networks used in precision agriculture." IEEE Transactions on Wireless Communications 15, no. 10 (2016): 6638-6650.

[88] Konstantopoulos, Christos, EftichiosKoutroulis, Nikolaos Mitianoudis, and AggelosBletsas. "Converting a plant to a battery and wireless sensor with scatter radio and ultra-low cost." IEEE Transactions on Instrumentation and Measurement 65, no. 2 (2015): 388-398.

[89] Santos, Ivairton Monteiro, and Carlos Eduardo Cugnasca. "Adaptive Strategies for Dynamic Setting of the Data Register Frequency in Wireless Sensor Networks." IEEE Latin America Transactions 12, no. 7 (2014): 1284-1291.

[90] Maheswararajah, Suhinthan, Saman K. Halgamuge, Kithsiri B. Dassanayake, and David Chapman. "Management of orphaned-nodes in wireless sensor networks for smart irrigation systems." IEEE transactions on signal processing 59, no. 10 (2011): 4909-4922.

[91] Kampianakis, Eleftherios, John Kimionis, Konstantinos Tountas, Christos Konstantopoulos, EftichiosKoutroulis, and AggelosBletsas. "Wireless environmental sensor networking with analog scatter radio and timer principles." IEEE Sensors Journal 14, no. 10 (2014): 3365-3376.

Optimal Power Allocation in Downlink Non-Orthogonal Multiple Access (NOMA)

Wajd Fahad Alghasmari[1], Laila Nassef[2]
Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah
Saudi Arabia

***Abstract*—Fifth generation of wireless cellular networks promise to enable better services anytime and anywhere. Non-orthogonal multiple access (NOMA) stands as a suitable multiple accessing scheme due to the ability to allow multiple users to share the same radio resource simultaneously via different domains (power, code, etc.). Through the introduced power domain, users multiplexed at the radio resource within different power levels. This paper studies power allocation in downlink NOMA, an optimization problem formulated that aims to maximize the system's sum rate. To solve the problem, a genetic algorithm based power allocation (GAPA) was proposed that uses genetic algorithm (GA) that employs heuristics to search for suitable solutions. The performance of the proposed power allocation algorithm compared with full search power allocation (FSPA) that gives an optimal performance. Results show that GAPA reaches a performance near to FSPA with lower complexity. In addition, GAPA simulated with various user paring algorithms. Channel state sorting based user pairing with GAPA achieves the best performance comparing to random user pairing algorithm and exhaustive user pairing.**

Keywords*—Non-orthogonal multiple access; power allocation; genetic algorithm; user pairing

I. INTRODUCTION

Conventional orthogonal multiple access (OMA) takes an incredible role starting from the first generation of wireless cellular networks (1G) to the fourth generation (4G) that uses orthogonal frequency division multiple access (OFDMA) for downlink network and single carrier (SC-OFDMA) for the uplink network [1]. Although OMA provides a various number of advantages to the network, it cannot handle massive connectivity under the diversity of quality of services (QoS) demands by users. Fairness, scarcity of the spectrum, and increasing number of connected devices add additional obstacles for OMA [2]. Non-orthogonal multiple access (NOMA) introduced for the fifth generation (5G) [3] [4]. NOMA represented as a suitable multiple accessing scheme due to the ability to handle an increased number of users and boosting the performance of the system such as spectrum efficiency [5]. Various schemes of non-orthogonal multiple access proposed such as Multiple User Shared Multiple Access (MUSA), Interleaver Division Multiple Access (IDMA), Spare Code Multiple Access (SCMA), and Pattern Division Multiple Access (PDMA). These schemes are divided into several categorize based on their properties [6], where MUSA and Resource Spread Multiple Access (RSMA) use spreading sequences. Otherwise, a structured coding matrix is used in SCMA and PDMA where IDMA based on interleaver and NOMA schemes is build based on the power domain.

In power domain NOMA, multiple users assigned to one resource block (RB) through different power levels utilizing superposition coding (SC) and successive interference cancelation (SIC) [7]. Channel condition plays a critical role in the performance and to the amount of power assigned to each user sharing the same RB [8], such that users coupled with distinctive channel conditions where the user with a bad channel condition allocated with higher power level than the user with good channel condition. One of the main designing issues to consider is the resource allocation in NOMA that can be identified by user pairing and power allocation. User pairing helps to identify the perfect couple of users to share a single RB while power allocation divide the power among the users sharing this RB. The optimal approach reached through the brute force strategy where all the possible solutions of user pairs and power allocation coefficients searched. Though the optimal performance gained, it is highly complex not to mention the complexity of SIC performance and excessive signaling overhead [9]. Reducing SIC execution can be established through the user pairing scheme that widely studied in different researches over the past years. Another designing aspect of downlink NOMA is power allocation which in addition helps boosting the performance of the system due the power domain multiplexing strategy. In [10], a general overview of downlink NOMA and a comparative simulation of different user pairing and power allocation schemes maintained. Random user pairing and channel state sorting based user pairing evaluated with fixed power allocation (FPA), through the simulation channel state sorting based user pairing achieves higher system sum rate than random user pairing. On the other hand, Full Search Power Allocation (FSPA) proven to reach optimal performance of the system though it is highly complex. Therefore, it is strictly essential to understand the tradeoff between the system performance and

the complexity of the system. Resource allocation received great attention from research society but optimal resource allocation still a very challenging task. In the downlink and uplink NOMA system, user clustering and power allocation were studied in [11]. Considering the channel gain difference a near-optimal user clustering algorithm proposed that aim to maximize the sum rate, for each cluster, an optimal power allocation strategy is given. The author in [12], propose iterative water-filling power allocation used with a greedy algorithm based user pairing to maximize users rate. Optimization of user pairing using matching theory proposed in [13], where a game between user and subchannel is estimated to match users into subchannels to maximize the sum rate of the system. Matching problem-based subchannel allocation and DC programming power allocation across subchannels and between users in each subchannel to boost energy efficiency studied in [14]. In [15], a bisection based iterative algorithm proposed for solving the non-convex problem yield from power allocation, the main objective of the proposed algorithm is to enhance the fairness of the system.

Genetic algorithm (GA) was introduced as a programming technique based on biological evolution [16]. It is characterized by overwhelming searching capability, which is usually utilized to find the optimal solution for complex problems. GA is employed for multiple domains of interest such as: data mining [17], fault diagnoses [18], cloud computing [19], Wireless Sensor Networks (WSN) [20], where cellular networks are no exception. In the LTE OFDMA system, GA utilized to learn antennas coverage pattern which leads to enhance the capacity of the system and decrease the network interference [21]. In the downlink NOMA system, a resource allocation algorithm using GA is proposed for pairing users that share the same frequency resource with an optimal power allocation strategy [22], results show that through the proposed algorithm a fast coverage to the target solution is achieved. On the other hand, GA utilized for power allocation in [23]. The proposed GA power allocation algorithm aims to maximize the achievable sum rate, results show that GA based NOMA overcome the performance of OMA. From the discussed works, reaching optimality in power allocation is still very challenging task especially with higher number of users sharing a single RB. Therefore, GA adopted for power allocation to maximize the system's sum rate considering power conditions and QoS of users which is assumed in this work as the minimum user's data rate.

The rest of the paper organized as follows: Section II discuss the mathematical model of the system. Power allocation problem formulated in Section III where the next Section IV presents the proposed power allocation algorithm. In Section V, the performance scenarios and performance metrics are evaluated to simulate, analyze, and compare performance. Finally, Section VI represents the conclusion and future works.

II. System Model

In a single cell, we study a downlink Multiuser NOMA with one Base Station (BS) and an arbitrary set of K users (kϵ 1,2,…, K) served over N RB (nϵ1,2,…,N). Channel gains of allocated users to nth RB ordered as $|h_{1,n}| \geqslant |h_{2,n}| \geqslant |h_{k,n}| \geqslant \cdots \geqslant |h_{K,n}|$, such that channel gain utilized to define the transmission power for each user in the users set allocated to that RB. Nevertheless, power allocation, user pairing, and SIC decoding order depend on channel gain sequence. At receiver side, the scheduled strong user on the nth RB uses SIC to exclude inter-user interference which is estimated through decoding other multiplexed signals of other users messages and subtracting these signals to be able to decode the signal of its own message. Weak users on other hand decode their own signals treating other signals as an interference, this process have negligible degradation on the performance due to power allocation policy followed in NOMA systems where weak users associated with high power levels. In downlink system, BS multiplex the messages of users sharing n^{th} RB via superposition coding, thus superimposed signal is expressed as:

$$x_n = \sum_{i=1}^{M} \sqrt{P_{i,n}} s_i \quad (1)$$

Considering the RB total power $\sum_{i=1}^{m} p_{i,n} = p_s$ where $p_{i,n}$ denote the power coefficient for UE_i in RB_n. The system total power is $\sum_{j=1}^{N} p_j = p_t$, where n and m (mϵ1,2,…, M) represent the index of RB and the index of users multiplexed over a RB, respectively. Assuming perfect knowledge of the channel state information of all users and Additive White Gaussian Noise (AWGN) channel is considered, user's k received signal on RB_n is calculated by:

$$y_{k,n} = \sqrt{P_{k,n}} |h_{k,n}| s_k + \sum_{i=1, i\neq k}^{M} \sqrt{P_{i,n}} |h_{k,n}| s_i + V_{k,n} \quad (2)$$

where the signals s_k and s_i is multiplexed over $|h_{k,n}|$ that represents the channel attenuation factor between user k and BS on RB_n. $V_{k,n}$ is the power spectral density with AWGN N_0 (W/Hz). Total bandwidth is divided equally among RB such that the bandwidth of a specific RB is defined as $B_n = \frac{B}{N}$. The total power on all RBs is assumed to be equivalent thus Signal Interference to Noise Ratio (SINR) for user K in RB_n is represented as [24]:

$$SINR_{K,n} = \frac{p_{K,n}|h_{K,n}|}{N_0 B + \sum_{i=1}^{K-1} p_{i,n}|h_{K,n}|} \quad (3)$$

UE_K as the weakest user allocated in RB_n do not perform SIC where the signals of other users allocated in the same RB and the environmental noise treated as an equivalent noise. In contrast UE_k, based on NOMA concept, performs SIC which enable successful decoding and subtracting (UE_{k+1}, UE_{k+2}, ….., UE_K) message signals on while treat other (UE_1, … ,UE_{k-2}, UE_{k-1}) message signals and the environmental noise as an equivalent noise. Thus SINR for user k in RB_n is expressed as:

$$SINR_{k,n} = \frac{p_{k,n}|h_{k,n}|}{N_0 B + \sum_{i=1}^{k-1} p_{i,n}|h_{k,n}|} \quad (4)$$

Assuming M=2, two users can concurrently share a RB_n. $R_{1,n}$ and $R_{2,n}$ represent user1 rate and user2 rate in RB_n and are calculated as follow:

$$R_{1,n} = B_n \log_2\left(1 + \frac{p_{1,n}h_{1,n}}{N_0 B_n}\right) \quad (5)$$

$$R_{2,n} = B_n \log_2\left(1 + \frac{p_{2,n}h_{2,n}}{p_{1,n}h_{2,n} + N_0 B_n}\right) \quad (6)$$

where $|h_{1,n}| \geqslant |h_{2,n}|$, $UE_{2,n}$ do not perform SIC. On the other hand, $UE_{1,n}$ needs to extract $UE_{2,n}$ signal then decode its own signal. Generally, under a successful decoding and no error propagation system with a randomized inter-cell interference that can be seen as white noise, where the power coefficient of $UE_{2,n}$ given higher ratio than $UE_{1,n}$ such as $p_1 \leqslant p_2$. Therefore, the achievable throughput of user k on RB_n is expressed as:

$$R_{k,n} = B_n \log_2\left(1 + \frac{p_{k,n}|h_{k,n}|}{N_0 B_n + \sum_{j=1}^{m-1} p_{j,n}|h_{k,n}|}\right) \quad (7)$$

Whereas, the total system sum rate equals to the summation of total sum rate calculated over the RBs, which is represented as:

$$R = \sum_{n=1}^{N} R_n = \sum_{n=1}^{N}\sum_{k=1}^{K} R_{k,n} \quad (8)$$

It is worthy to mention that not only exclusive sum rate optimization of the system is provided, a spectral efficiency and energy efficiency performance metrics is also simulated. For each RB in NOMA system given the total sum rate calculated at the RB_n as R_n and the bandwidth of this RB_n is B_n, then the spectral efficiency of the regarded RB_n is expressed as:

$$SE_n = \frac{R_n}{B_n} \quad (9)$$

Thus, the total spectral efficiency or as regarded as the system's spectral efficiency is calculated by:

$$SE = \sum_{n=1}^{N} SE_n \quad (10)$$

Additionally, energy efficiency over RB_n given that p_s and p_c as the total RB power and the additional circuit power consumption, respectively. Therefore, energy efficiency of the RB_n is defined as:

$$EE_n = \frac{R_n}{p_s + p_c} \quad (11)$$

where the total energy efficiency or as known by system's energy efficiency is expressed as:

$$EE = \sum_{n=1}^{N} EE_n \quad (12)$$

Due to the effect of the factor of dividing the power to paired users, an optimization of power allocation is needed. Additionally, most of the works presented to optimize power allocation consider two user multiplexing strategy according to simplicity, rather multiple user multiplexing strategy need to be researched. Motivated by that, the main contribution of this work is to propose a power allocation algorithm that maximize system's sum rate, the problem related to the power allocation formulated in the next section followed by the solution approach.

III. Problem Statement

In this section, we formulate a maximization of system's sum rate power allocation as an optimization problem. In order to enhance system's sum rate, it is important for each user to reach or exceed a minimum rate. Therefore, power allocation optimization problem is formulated as:

$$\max_{p_{m,n}} \sum_{n=1}^{N} R_n$$

Subject to:

C1: $R_{k,n} \geq R_{min}$

C2: $\sum_{m=1}^{M} p_{m,n} = p_s$

C3: $p_{m,n} > 0$

Where R_{min} represents the minimum user rate requirement, constraint C1 applies that each user data rate have to reach or exceed a specific user rate. Power allocation constraints discussed in C2 and C3, where the available power of an individual RB divided among the users sharing it as expressed in C2 such that the summation of power coefficients equals to total power of the RB. C3 on other hand indicates that power allocation coefficient must be higher than zero.

IV. Genetic Algorithm based Power Allocation (GAPA)

Genetic algorithm (GA) adopted as an intelligent search algorithm to find the optimal solution, is defined as an optimization method that explores huge search space based on a powerful meat-heuristic. To solve the power allocation problem formulated in the previous section, the Genetic Algorithm based Power Allocation (GAPA) processes are shown in Algorithm 1. The process of resource allocation performed iteratively beginning by pairing users to a specific RB then GAPA is utilized. Therefore, the processes of GAPA held where the number of genes in chromosomes is highly dependent on the multiplexing number such that for two users sharing the same RB (scenario1), the genes of chromosomes in the generations is equivalent to 2. Moreover, considering the case of three users multiplexing genes of chromosomes in every generation is equal to 3 also. Due to the nature of solutions, a string of real number representation was adopted such that the power level of an individual user presented in real numbers.

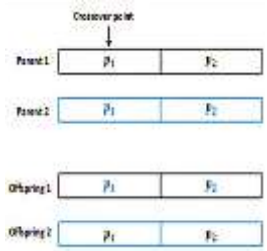
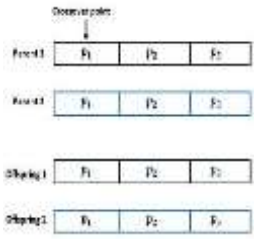
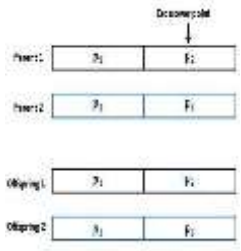
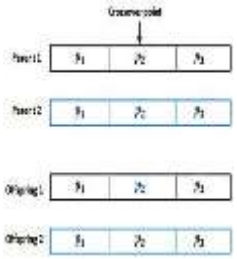
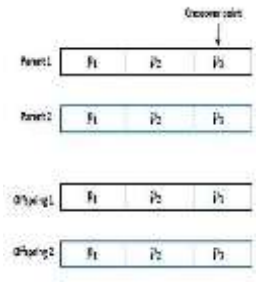
A random set of L chromosomes are generated, each representing a feasible solution. For some cases, the coverage of the population was maintained without a limitation in size. Rather in this work, we limit the number of chromosomes in a generation to L= 100 and the number of generations to be 50. Chromosomes of each generation including the initial generation evaluated based on a predefined fitness function represent the optimization objective, by that means the fitness function in this paper is the sum rate formulated in (8). Then generations of chromosomes go through other GA operations of selection, crossover, and mutation to explore and invoke modified solutions.

Algorithm 1: GAPA

```
for RB1 to RBn do
    for l=1 to L do
        //generate the initial population of chromosomes
        //chromosomes represent the power coefficient for users
          sharing the RB power= (P1,n ,..., Pm,n);
        for g=1 to G do
            //evaluate each chromosome through fitness function
              calculated in (4.8)
            //create a measurement array that calculate not only the
              fitness function additionally rates of users sharing RB
              (4.7), sum rate (4.8), spectral efficiency (4.9) and
              energy efficiency (4.11)
            Measurement=(R1,n ,...,Rm,n,Rn,SEn,EEn);
            for c=1 to L*crossoverratio do
                //perform crossover over a portion of population
                //randomly select parents from the population and
                  swap the genes between every two parents
                if crossoverpoint=1 then
                    swap all genes between the parents;
                    else if crossoverpoint=2 then
                        swap the genes from the second gene between the
                        parents;
                        else
                            swap the third gene between the parents;
                        end
                    end
                end
            end
            for m=1 to L*mutationratio do
                //perform mutation over a portion of the population
                if mutationpoint=1 then
                    change the value of the first gene randomly;
                    else if mutationpoint=2 then
                        change the value of the second gene randomly;
                        else
                            swap the third gene between the parents;
                        end
                    end
                end
            end
            Repeat evaluation;
            for l=1 to L do
                select the next population based on the roulette wheel selector;
            end
        end
    end
end
```

Through crossover and mutation, or as known by reproduction, a new offspring of chromosomes are generated where crossover ratio is assumed to be 0.5 and mutation ratio equals to 0.25. A 50% of the current population perform crossover where two arbitrary selected chromosomes referred to as parents swap their genes such that in GAPA, crossover have to consider the main requirements such that after swapping the genes among two chromosomes power levels of users with higher channel gain should not be higher than the power level of users with lower channel gain. Another point to be taken into consideration that the summation of power levels after exchanging the genes should be equal to the total RB power based on the second condition of the formulated problem, where a one point crossover is applied. The process of crossover is based on the number of genes summarized in Table I.

TABLE I. CROSSOVER OPERATION IN GAPA

Crossover cases	**Two user multiplexing (scenario 1) M=2**	**Three user multiplexing (scenario 2) M=3**
Case 1		
Case 2		
Case 3		

On the other hand, 25% of chromosomes in a population chosen for mutation where random chromosomes are selected such that for the randomly chosen gene the value will be altered. Based on the dependability of the solved problem to real numbers a uniform mutation is utilized. Therefore, the altered value on a specific gene is randomly generated such that it should not exceed an upper bound or not be less than a lower bound which in our case determined by the total

transmission power on a RB such that by adding the power ratios assigned to each user associated in this RB the value should be equal to the total RB power. Selection then held which plays a huge role in forming the new generations, from different approaches for selection a roulette wheel selection is employed which is based on a probability distribution that gives the probability of selection to each chromosome. Thus, the probability of selection for chromosome j is defined as:

$$B_j = \frac{R_j}{\sum_{i=1}^{L} R_i} \quad (13)$$

These processes go in a cycle until the stopping criteria is reached which is assumed to be the number of generations that assumed earlier to be 50. Therefore, user pairing and GAPA performed iteratively for each RB available in the system.

V. Results and Discussion

For the simulation results, the performance of our proposed power allocation algorithm is compared with FSPA that present an optimal solution. The simulation runs into a NOMA system with one BS. A single transmitting antenna and multiple users, each is occupied with one receiver antenna in a single cell. System's sum rate, spectral efficiency, and energy efficiency are evaluated for GAPA with exhaustive search user pairing in downlink NOMA system. Through this simulation, we assume the channel to be the product of free-space path loss and Gaussian white noise. Table II presents the most commonly used simulation parameters that is used commonly in similar researches. RBs are characterized with equivalent amount of bandwidth and downlink transmission power. For this scenario, the number of users in the system is considered to be 6 and 12 due to the high complexity produced by the large number of users in the system. We set the minimum rate that can be achieved by each user R_{min} to 100 kbps. In addition the number of GA individuals, crossover ratio, mutation ratio, and the number of generations as a stop criteria are set to 100, 0.50, 0.25, 50, respectively.

In the beginning, the performance of the system with different number of users served among the cell is simulated. Therefore, the number of users are assumed to be either K=6 or K=12. Both cases are evaluated with GAPA and FSPA for power allocation and exhaustive search user pairing, with three users multiplexed per RB (M=3). That implies with K=6, the number of RBs needed for users to be served among equals N=2. On other hand, with K=12 the number of RBs for users to be paired on as a triple is N=4.

A comparison between FSPA and GAPA as a function of the system's sum rate is illustrated in Fig. 1. The figure shows that with the increase in the transmitted power, the gap between FSPA and GAPA increase for both downlink NOMA systems. Considering the system with six users (K=6), both power allocation algorithms reaching the same performance at lower transmitted power while GAPA outperforms FSPA for higher transmitted power due to the heuristic searching method. On the other hand, in downlink NOMA with twelve users (K=12) GAPA achieves superior system's sum rate than FSPA for higher transmitted power where on lower transmitted power FSPA is better though GAPA is low complex.

TABLE II. Simulation Parameters

Parameter	Value
Transmitted Power (p_t)	316 mW (25 dBm)
Total Bandwidth (B)	5 MHz
Number of resource blocks RPs (N)	24
Number of subcarrier	12 per RP
Noise Spectral Density (N_0)	-150 dBw/Hz
Channel estimation	Ideal
Channel	AWGN
Traffic Model	Full Buffer
Circuit power (p_c)	1 w
Multiplexed users over single RB (M)	2,3
Number of users (K)	6,12
Number of transmit antenna at BS	1
Number of receiver antenna at UE	1
Minimum user's data rate (R_{min})	100 kbps
Number of chromosomes per generation (L)	100
Number of generations (G)	50
Crossover ratio	0.50
Mutation ratio	0.25

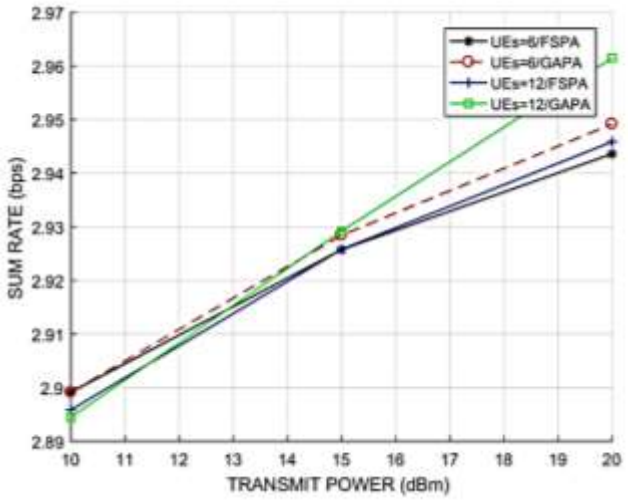


Fig. 1. Sum Rate Performance of 3-users Multiplexing Downlink NOMA (M=3) with FSPA and GAPA.

Fig. 2 and Fig. 3 illustrate the performance of the system based on spectral efficiency that is expressed as the ratio of sum rate and bandwidth. In downlink NOMA with K=6, GAPA performs better than FSPA especially with higher transmitted power as shown in Fig. 2. Additionally, Fig. 3 illustrates the performance of spectral efficiency in downlink NOMA system with K=12. Searching based on heuristics benefit with the system's performance such that GAPA is more spectrally efficient than FSPA. Number of users served in the system effect the performance where with the growth in the number of users the spectral efficiency increase.

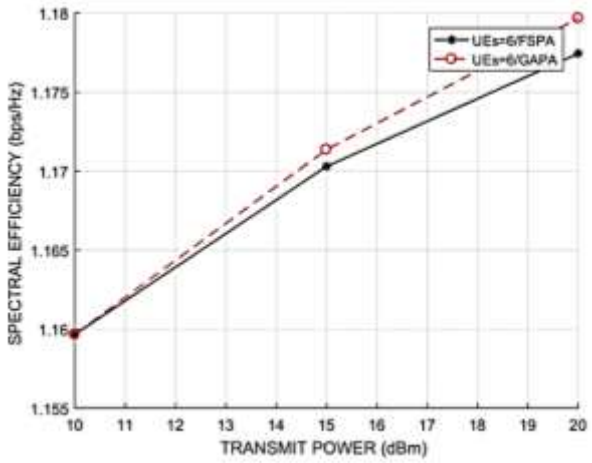


Fig. 2. Spectral Effeciency Performance of 6-users (K=6) Downlink NOMA with FSPA and GAPA.

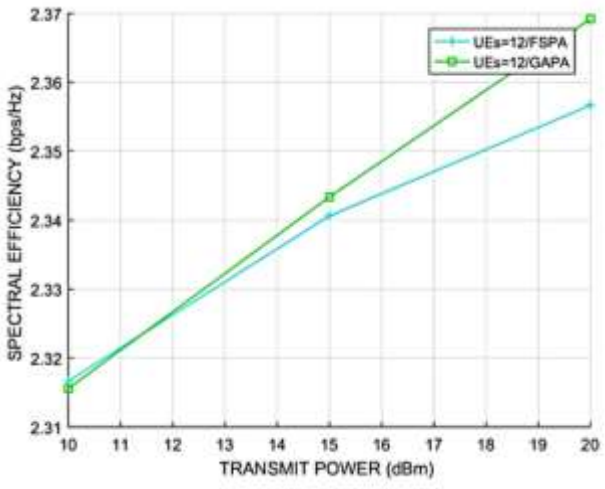


Fig. 3. Spectral Effeciency Performance of 12-users (K=12) Downlink NOMA with FSPA and GAPA.

Energy efficiency (EE) of the system is presented in Fig. 4. The value of P_c is assumed to be constant and equal to 1W. Results shows that both FSPA and GAPA have reached an equivalent performance based on overall system energy efficiency for 6 users case. In contrast, downlink NOMA with K=12 GAPA has consumed less energy than FSPA due to the utilized searching technique. It is also found that with the increase in transmission power system, the energy efficiency decreased due to transmission with large power to maximize overall sum rate. Referring to Fig. 1, GAPA with 12 users gain has the highest system sum rate which lead to deceased system EE.

Time complexity of simulated power allocation algorithms in both systems is evaluated. Fig. 5 shows that GAPA in both system decreases the complexity of the system with higher performance gain than FSPA that represent the best optimum performance. FSPA used in downlink NOMA with K=12 takes the longest time where with GAPA the time complexity decreased significantly. Additionally, in downlink NOMA with K=6 FSAP reach higher complexity in GAPA.

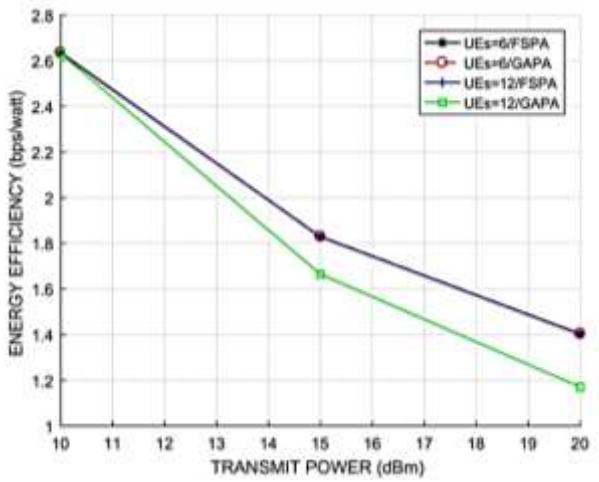


Fig. 4. Energy Effeciency Performance of Downlink NOMA with FSPA and GAPA.

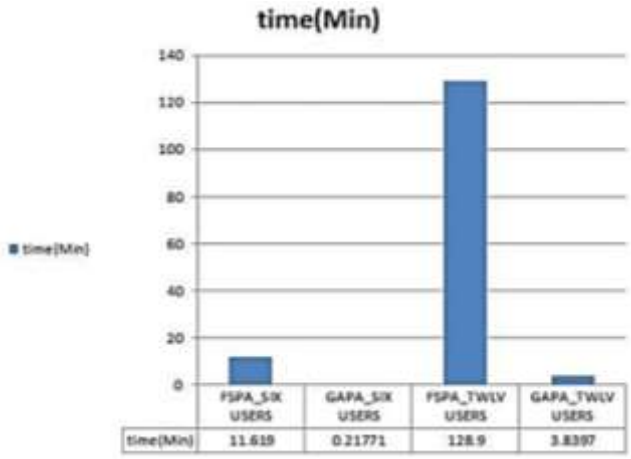


Fig. 5. Time Complexity of Power Allocation Algorithm in Downlink NOMA Systems.

A second scenario for simulation is done through a simulation of different resource allocation schemes, where we assume that these schemes differentiated by the user pairing algorithms. Three user pairing schemes simulated with GAPA. A random user pairing that works by pairing arbitrary users into a cluster to be allocated into an individual RB, a channel state based sorting user pairing that is based on dividing the users based on their channel conditions into different NOMA clusters each must contain a strong and weak users. In addition, exhaustive search based user pairing is considered. All algorithms invoked with GAPA.

As illustrated in Fig. 6 to Fig. 8, channel state information based sorting user pairing gain the best performance among other algorithms in term of system sum rate, spectral efficiency and energy efficiency. Such that with the increase in transmit power the performance gain of the three user pairing algorithms increase as well. System's sum rate and overall spectral efficiency performance of exhaustive approach are slightly higher than random user pairing, where in term of energy efficiency the performance of these two user pairing algorithms are equivalent.

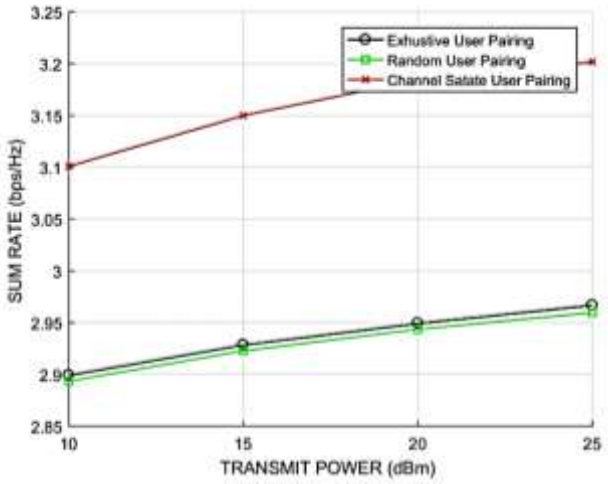


Fig. 6. Sum Rate Performance of 3-users Multiplexing Downlink NOMA (M=3) with Different User Pairing Algorithms.

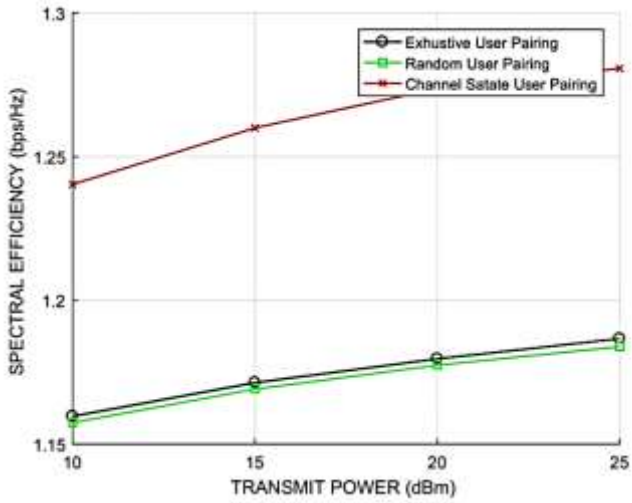


Fig. 7. Spectral Efficiency Performance of 3-users Multiplexing Downlink NOMA (M=3) with Different User Pairing Algorithms.

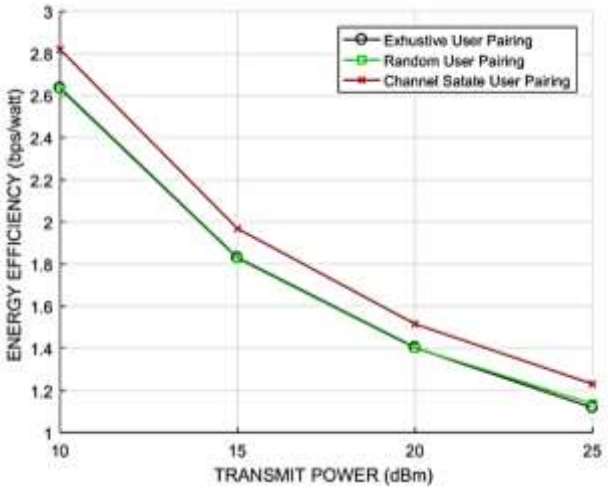


Fig. 8. Energy Efficiency Performance of 3-users Multiplexing Downlink NOMA (M=3) with Different User Pairing Algorithms.

Finally, in downlink NOMA with K=12 the effect of the number of users multiplexed over the same radio resource is investigated. For evaluation, two multiplexing scheme M=2 and three multiplexing scheme M=3 for downlink NOMA system is simulated. Therefore, the number of RB needed for tow multiplexing scheme M=2 given as N=6. Where on the case with three user multiplexing scheme M=3, four RBs N=4 needed. The result in Fig. 9 shows that the sum rate of the system with three users multiplexing system M=3 overcome system's sum rate with two user multiplexing scheme M=2. Therefore, increasing the number of users sharing the same resource helps to increase system's sum rate performance. In addition, utilization of GAPA with both multiplexing schemes scenarios increase the performance of the system especially with higher transmitted power levels.

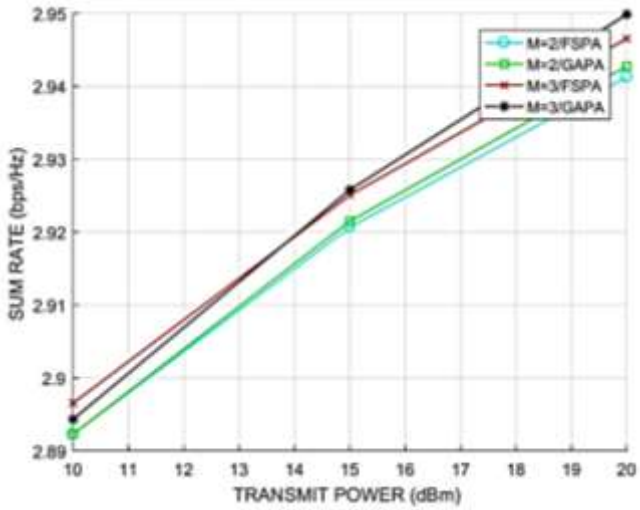


Fig. 9. Sum Rate Performance of 2-users and 3-users Multiplexing Downlink NOMA (M=2 and M=3).

VI. Conclusion

Next generation cellular networks needs a new and highly effective technologies to be adopted, one of the promising recommended technology for radio accessing technique is NOMA. The basic concept of NOMA is to utilize power as its new diminution. In a downlink NOMA system, power allocation optimization problem is formulated where total subchannel power and minimum user's data rate are taken in consideration. Genetic algorithm power allocation (GAPA) is proposed to solve the problem, which can not only achieve high performance gain but can decrease the complexity. GAPA is utilized with an exhaustive user pairing scheme and is evaluated through comparing the performance with FSPA. Results show that the proposed algorithm reach a good performance in addition the complexity is decreased. GAPA outperform FSPA with large number of users and high transmission power. Moreover, exhaustive, random, and channel state based sorting user pairing algorithms were invoked with GAPA. Through these experiments, channel state based user pairing overcome the other two user pairing algorithms based on system's sum rate, spectral efficiency, and energy efficiency. Finally, the impact of multiplexed users in a single subchannel was studied, where results revealed that with

higher number of multiplexed users, the system's sum rate increased. For future investigations, enabling the proposed algorithm with a downlink system with imperfect channel state information and further enhancement of the spectral efficiency by adding MIMO technology needed to be studied.

REFERENCES

[1] Kalhoro, S., Umrani, F. A., Khanzada, M. A., & Rahoo, L. A. (2019). Matched Filter Based Spectrum Sensing Technique for 4G Cellular Network. Mehran University Research Journal of Engineering and Technology, 38, 973–978. https://doi.org/https://doi.org/10.22581/muet1982.1904.10.

[2] Vamvakas, P., Tsiropoulou, E. E., & Papavassiliou, S. (2019). Dynamic Spectrum Management in 5G Wireless Networks: A Real-Life Modeling Approach. In IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. Paris, France,: IEEE. https://doi.org/10.1109/INFOCOM.2019.8737443.

[3] Vaezi, M., & Poor, H. V. (2019). NOMA: An Information-Theoretic Perspective. In Multiple Access Techniques for 5G Wireless Networks and Beyond (pp. 167–193). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-92090-0_5.

[4] Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., Li, A., & Higuchi, K. (2013). Non-orthogonal multiple access (NOMA) for cellular future radio access. In 2013 IEEE 77th Vehicular Technology Conference (VTC Spring) (pp. 1–5). Dresden. https://doi.org/10.1109/VTCSpring.2013.6692652.

[5] Agyapong, P. K., Iwamura, M., Staehle, D., Kiess, W., & Benjebbour, A. (2014). Design considerations for a 5G network architecture. IEEE Communications Magazine, 52(11), 65–75. https://doi.org/10.1109/MCOM.2014.6957145.

[6] Liu, Y., Qin, Z., Elkashlan, M., Ding, Z., Nallanathan, A., & Hanzo, L. (2017). Non-orthogonal multiple access for 5G and beyond. Proceedings of the IEEE, 105(12), 2347–2381. https://doi.org/10.1109/JPROC.2017.2768666.

[7] Choi, J. (2018). Throughput analysis for multiuser diversity of two users with SIC in NOMA systems. In 2018 International Conference on Signals and Systems (ICSigSys) (pp. 120–124). Bali. https://doi.org/10.1109/ICSIGSYS.2018.8372649.

[8] Ding, Z., Peng, M., & Poor, H. V. (2015). Cooperative Non-Orthogonal Multiple Access in 5G Systems. IEEE Communications Letters, 19(8), 1462–1465. https://doi.org/10.1109/LCOMM.2015.2441064.

[9] Benjebbour, A., Saito, Y., Kishiyama, Y., Li, A., Harada, A., & Nakamura, T. (2013). Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access. In 2013 International Symposium on Intelligent Signal Processing and Communication Systems (pp. 770–774). Naha. https://doi.org/10.1109/ISPACS.2013.6704653.

[10] Alghasmari, W. F., & Nassef, L. (2020). Power Allocation Evaluation for Downlink Non-Orthogonal Multiple Access (NOMA). International Journal of Advanced Computer Science and Applications(IJACSA), 11(4). https://doi.org/10.14569/IJACSA.2020.0110417.

[11] Ali, M. S., Tabassum, H., & Hossain, E. (2016). Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems. IEEE Access, 4, 6325–6343. https://doi.org/10.1109/ACCESS.2016.2604821.

[12] Cai, W., Chen, C., Bai, L., Bai, Y., & Choi, J. (2017). Subcarrier and power allocation scheme for downlink OFDM-NOMA systems. IET Signal Processing, 11(1). https://doi.org/10.1049/iet-spr.2016.0188.

[13] Di, B., Bayat, S., Song, L., & Li, Y. (2015). Radio Resource Allocation for Downlink Non-Orthogonal Multiple Access (NOMA) Networks Using Matching Theory. In 2015 IEEE Global Communications Conference (GLOBECOM). San Diego, CA, USA: IEEE. https://doi.org/10.1109/GLOCOM.2015.7417643.

[14] Fang, F., Zhang, H., Cheng, J., & Leung, V. C. M. (2016). Energy-Efficient Resource Allocation for Downlink Non-Orthogonal Multiple Access Network. IEEE Transactions on Communications, 64(9), 3722–3732. https://doi.org/10.1109/TCOMM.2016.2594759.

[15] Timotheou, S., & Krikidis, I. (2015). Fairness for Non-Orthogonal Multiple Access in 5G Systems. IEEE Signal Processing Letters, 22(10), 1647–1651. https://doi.org/10.1109/LSP.2015.2417119.

[16] Kramer, O. (2017). Genetic Algorithm Essentials. Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-52156-5.

[17] Ting, C.-K., Wang, T.-C., Liaw, R.-T., & Hong, T.-P. (2017). Genetic algorithm with a structure-based representation for genetic-fuzzy data mining. Soft Computing, 21, 2871–2882. https://doi.org/https://doi.org/10.1007/s00500-016-2266-z.

[18] Zhu, X., Xiong, J., & Liang, Q. (2018). Fault Diagnosis of Rotation Machinery Based on Support Vector Machine Optimized by Quantum Genetic Algorithm. IEEE Access, 6, 33583–33588. https://doi.org/10.1109/ACCESS.2018.2789933.

[19] Mahmood, A., Khan, S. A., & 3, R. A. B. (2017). Hard Real-Time Task Scheduling in Cloud Computing Using an Adaptive Genetic Algorithm. Computers 2018, 6(2), 15. https://doi.org/https://doi.org/10.3390/computers6020015.

[20] Jha, S. K., & Eyong, E. M. (2018). An energy optimization in wireless sensor networks by using genetic algorithm. Telecommunication Systems, 67, 113–121. https://doi.org/https://doi.org/10.1007/s11235-017-0324-1.

[21] Yang, X., Wang, Y., Zhang, D., & Cuthbert, L. (2010). Resource Allocation in LTE OFDMA Systems Using Genetic Algorithm and Semi-Smart Antennas. In 2010 IEEE Wireless Communication and Networking Conference. Sydney,Australia: IEEE. https://doi.org/10.1109/WCNC.2010.5506423.

[22] Gemici, Ö. F., Kara, F., Hokelek, I., Kurt, G. K., & Çırpan, H. A. (2017). Resource allocation for NOMA downlink systems: Genetic algorithm approach. In 2017 40th International Conference on Telecommunications and Signal Processing (TSP). Barcelona, Spain: IEEE.

[23] Ma, X., Wu, J., Zhang, Z., Zhang, Z., Wang, X., Chai, X., … Dai, X. (2017). Power Allocation for Downlink of Non-orthogonal Multiple Access System via Genetic Algorithm. In 5GWN 2017 (pp. 459–470). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-72823-0_43.

[24] Hanif, M. F., Ding, Z., Ratnarajah, T., & Karagiannidis, G. K. (2016). A Minorization-Maximization Method for Optimizing Sum Rate in the Downlink of Non-Orthogonal Multiple Access Systems. IEEE Transactions on Signal Processing, 64(1), 76–88. https://doi.org/10.1109/TSP.2015.2480042.

Comparing the Accuracy and Developed Models for Predicting the Confrontation Naming of the Elderly in South Korea using Weighted Random Forest, Random Forest, and Support Vector Regression

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

***Abstract*—Since dementia patients clearly show the retrogression of linguistic ability from the early stage, evaluating cognitive and language abilities is very important when diagnosing dementia. Among them, naming is an essential item (sub-test) that is always included in the dementia-screening test. This study developed confrontation naming prediction models using support vector regression (SVR), random forest, and weighted random forest for the elderly in the community and identified an algorithm showing the best performance by comparing the accuracy of the models. This study used 485 elderly subjects (248 men and 237 women) living in Seoul and Incheon who were 74 years old or older. Prediction models were developed using SVR, random forest, and weighted random forest algorithms. This study revealed that the root mean squared error of weighted random forests was the lowest when comparing the prediction performance using models based on SVR, random forest, and weighted random forest. Future studies are needed to compare the prediction performance of weighted random forest with other machine learning models by calculating various performance indices such as sensitivity, specificity, and harmonic mean using data from various fields to prove the superior prediction performance of weighted random forest.**

Keywords*—Confrontation naming; generative naming; support vector regression; random forest; weighted random forest

I. Introduction

The elderly population is rapidly increasing worldwide as the life expectancy is extended because the socioeconomic level has been improved and medical science has been advanced. In particular, aging is progressing faster in South Korea than in Europe, the United States, and Australia since South Korea has experienced an increase in the elderly population and a low birth rate at the same time. South Korea entered an aged society in 2017 with the proportion of the elderly population (65 years old or older) more than 14% [1]. It is also forecasted that South Korea will enter a super-aged society in 2026, indicating that the proportion of the elderly population will exceed 20% in 2026 [1]. When the elderly population increases, and the occurrence of senile diseases also increases. Particularly, the incidence of dementia has rapidly increased and it was forecasted that it would reach 633,000 in 2020, a large increase from 220,000 in 2010 [2]. As the number of patients with dementia increases, geriatric medicine has been actively studied the characteristics of early dementia and the early detection of dementia [3,4,5].

Communication abilities, as well as cognitive abilities such as memory, are deteriorated distinctively in the aging process. Kang et al. (2001)[6] reported that 41.4% of the elderly population in South Korea experienced several difficulties in communication during daily life activities. As aging progresses, the elderly gradually have more difficulties in understanding and expressing language [7,8], and also experience difficulties in inference and reminiscence [9]. Particularly, previous studies [10,11], which evaluated the linguistic performance of healthy elderly people, revealed that the elderly had an inferior generative naming ability, indicating the ability to freely recall words, to young adults.

Recently, confrontation naming has drawn attention as an effective differentiation indicator of senile cognitive disorders such as dementia. Since patients with dementia clearly show the retrogression of linguistic ability from the early stage, evaluating cognitive and language abilities is essential when diagnosing dementia [12,13,14]. Among them, naming is an essential item (sub-test) that is always included in the dementia screening test. It has been forecasted that the number of dementia patients will increase as the proportion of the elderly population increases. Therefore, accurately understanding the risk factors of cognitive disorders, diagnosing them early, and providing appropriate rehabilitation accordingly are a crucial issue in the field of geriatrics and gerontology [15].

Over the past decade, supervised learning-based machine learning algorithms such as support vector regression (SVR), weighted random forest, and random forest have been widely used as a way to identify complex risk factors of diseases [16,17,18]. Although ensemble machines have been reported to have better prediction performance in classifying binary data such as the presence or absence of diseases compared to decision trees such as classification and regression trees [19,20,21], most studies used regression models and decision trees to predict the cognitive disorders in old age by using demographic and other factors [22,23], and only a few studies have used ensemble machines. In addition, as far as we are aware, no study has attempted to predicting the communication characteristics of healthy South Korean elderly people in the normal aging process using an ensemble machine. This study

developed confrontation naming prediction models using SVR, random forest, and weighted random forest for the elderly in the community and identified an algorithm showing the best performance by comparing the accuracy of the models.

II. Research and Methods

A. Subjects

This study used 485 elderly subjects (248 men and 237 women) living in Seoul and Incheon who were 74 years old or older. Selection criteria were (1) those without a history of neurological diseases such as stroke or Parkinson's disease, (2) those who received 24 or higher points from the Korean version of Mini-Mental State Exam (K-MMSE) and fell within the normal range, (3) the elderly who did not have visual and hearing impairment for conducting the study, and (4) the elderly who did not have depression according to the results of the Korea-Geriatric Depression Scale Short form (K-GDS-S). Power analysis was conducted using G-Power version 3.1.9.7 (Universität Mannheim, Mannheim, Germany) (Fig. 1). When predictor variables were nine, alpha=0.05, power (1-B) =0.95, and effect size (f^2) was 0.25, the number of samples was 400, indicating that the sample size of this study exceeded the appropriate sample size to conduct statistical tests (Fig. 2).

B. Definition of Measurements and Variables

This study measured the confrontation naming ability by using the Short forms of the Korean-Boston Naming Test (K-BNT-15) because the elderly have limited attention ability and it is difficult to conduct an examination for a long time. [24]. K-BNT-15 is a task to evaluate the confrontation naming ability by looking at the presented picture and saying the name of it. It gives one point per correct answer, and the total score was 15 points. The cut-off score is eight points [24].

Executive function, visuospatial ability, memory, attention concentration, language function, and orientation were measured using the Korean Version of Montreal Cognitive Assessment (K-MoCA) [25]. K-MoCA is a standardized cognitive screening test that can effectively discriminate various dementia patients including mild cognitive impairment (MCI) and vascular dementia. It is composed of multiple aspects of executive functions (4 points; trail-making B task, a phonemic fluency task, and a verbal abstraction task), visuospatial abilities (4 points; a three-dimensional cube copy and a clock-drawing task), memory (5 points; the short-term memory recall task), sustained attention task (6 points; number memorization, target detection using tapping, and subtracting by 7 from 100), language (5 points), and orientation (6 points), and the total score of it is 30 points.

Generative naming was measured using both semantic fluency test and phonetic fluency test among the items of Controlled OralWord Association Test (COWAT), a sub-test of Seoul Neuropsychological Screening Battery (SNSB)[26]. The semantic fluency task requires the activation of lexical-semantic, and the subject was asked to speak the vocabulary within the "animal" category for one minute. The phonetic fluency test requires the activation of the phonetic-lexical network, and this study conducted only the "k" phoneme. The examiner recorded all responses spoken by the subject for one minute in order on the response sheet, and the correct responses were calculated by counting the total number of words.

Picture description was measured using the task of observing and describing "seashore", an item in the self for oneself section of the Korean version of the Western Aphasia Battery (K-WAB)[27]. This study calculated the correct information unit (CIU ratio, %) according to Eq. 1, indicating the proportion of words providing appropriate and correct information among the descriptions of the "seashore".

$$\text{CIU ratio (\%)} = \text{Number of CIUs} / \text{Total Number of Words} \times 100 \quad (1)$$

Working memory was measured using the Digit Span test, a subtest of the Korean Wechsler Adult Intelligence Scale (K-WAIS) [28]. The Digit Span is measured by repeating forward or backward the numbers called by an examiner and it reflects working memory. Digit span-forward starts with 3 numbers, and the number of numbers to be memorized increases by one in the next step. The last seventh step has nine numbers to be memorized. Each step has two trials, and the second trial is conducted only when the subject fails in the first trial. It was scored by recording the number of digits of the step accurately performed by the subject, and the total score is 14 points. Digit span-backward is a task to listen to a series of numbers and repeat the numbers in reverse order, and it was conducted and scored in the same way as the digit span-forward.

Depression was measured using the Short form of Geriatric Depression Scale (SGDS). Sheikh & Yesavage (1986)[29] developed the SGDS based on diagnostic validity studies on the existing Geriatric Depression Scale (GDS). They selected 15 items showing the highest correlation with depression out of the 30 items of the GDS. At the time of development, they reported that the correlation coefficient (r) between the GDS and the SGDS was 0.84, indicating a strong correlation. The cut-off score defining depression was set as 6 points based on the results of previous studies [30,31].

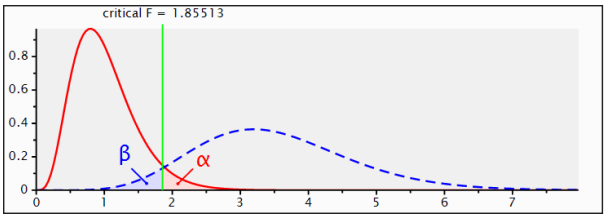


Fig. 1. Result of Power Analysis using G-Power.

Analysis:	A priori: Compute required sample size			
Input:	Effect size f	=	0.25	
	α err prob	=	0.05	
	Power (1-β err prob)	=		0.95
	Numerator df	=	10	
	Number of groups	=	2	
	Number of covariates	=		9
Output:	Noncentrality parameter ë	=		25.0000000
	Critical F =	1.8550636		
	Denominator df	=	389	
	Total sample size	=	400	
	Actual power	=	0.9503386	

Fig. 2. Results of Estimating the Number of Samples Needed for Statistical Analyses.

Explanatory variables were age, gender (male, female), educational level (middle school graduate and below and middle school graduate above), mean monthly household income (<1.5 million KRW, ⩾1.5 million KRW and < 2.5 million KRW, and ⩾2.5 million KRW), living together or not after marriage (living with a spouse, bereavement, and separation from a spouse), smoking (non-smoking and smoking), drinking (non-drinking and drinking), working memory (total score), pictures description (CIU ratio), prevalence of depression, generative naming (total score), executive function (total score), visuospatial ability, memory (total score), attention concentration (total score), language function (total score), and orientation (total score). Table I shows the results of descriptive statistics on the general characteristics of the subjects.

C. Development of a Confrontation Naming Prediction Model for Elderly People in South Korea

SVR is a regression model based on a support vector machine (SVM). SVR is an extension of SVM, so that it can be applied to regression analysis [32]. It is used to predict a random loss value by introducing an e-insensitive loss function [32]. SVR has the advantage of having high explanatory power even for data with nonlinearity or complex patterns. On the other hand, it also has the disadvantage that it requires a long learning time due to high computational complexity and it is difficult to interpret the model because it is impossible to analyze the direct relationship between the independent variable and the dependent variable. Moreover, SVR converts a nonlinear feature space that cannot be separated linearly into a high-dimensional linear regression problem by using a kernel function for nonlinear expansion. Linear, polynomial, and radial basis kernel functions are generally used for this process. The concept of SVR is presented in Fig. 3.

Random forest is one of the ensemble techniques that generate multiple tree models using bootstrap samples and predict the outcome by synthesizing the models. Random forest does not use all p-dimensional explanatory variables, but it splits tree by randomly selecting m-dimensional explanatory variables smaller than that. Random forest has the advantage of being able to use out of bag (OOB) samples because it uses bootstrap samples [34,35]. The importance of the variable can be easily calculated through permutation, and the mean square error (MSE) of the OOB sample is calculated using the regression tree model generated by the bootstrap samples. The concept of random forest is presented in Fig. 4.

Weighted random forest is one of the ensemble techniques that conducts model averaging by applying the same weight to each tree model. Since random forest generated by bootstrapping, there is a possibility that the random forest is composed of models showing good performance and those showing bad performance. If the model averaging is performed with giving more weight to good tree models, it can provide better prediction power than the existing random forest models giving equal weight. Weighted random forest algorithm was developed based on this concept (Fig. 5). Weighted random forest also uses OOB samples as random forest does. Regarding b = 1, …, B, when the MSE e(b) of an OOB sample O(b) was calculated with the tree model Tr(fb), generated with the b^{th} bootstrap sample θ(b), it is assumed that a model with a large e(b) is a bad tree model and a model with a small e(b) is a good tree model. A model averaging technique using a weight given to each tree model (Tr(fb)) by using the calculated e(b) is defined as the weighted random forest. This model used Akaike weights [37] for selecting AIC models.

D. Evaluating the Prediction Performance of Machine Learning Models

Multiple linear regression analysis builds models by applying a regression coefficient estimation method using the least squares method. Random forest limited the number of developed decision tree models to 100. SVR was analyzed using the linear kernel function, the most basic kernel function. It was analyzed by setting c (a parameter determining the generalization of the regression model) as 15.0 and e-insensitive loss function (a precision parameter) as 0.001. This study compared the root mean squared error (%) of developed models to compare their prediction performance. Since random forest has randomness, and the random seed was fixed to seed No. 123789 while reiterating the models.

TABLE I. RESULTS OF DESCRIPTIVE STATISTICS ON THE GENERAL CHARACTERISTICS OF THE SUBJECTS

Factor	Mean±SD	Minimum	Maximum
Age, year	69.3±8.3	60	83
Education, year	10.1±3.1	6	16
Executive function	2.2±1.4	0	4
Visuospatial ability	3.5±0.7	1	4
Working memory (digit span: forward)	6.2±2.2	3	12
Working memory (digit span: backward)	4.6±2.1	1	11
Memory	2.1±1.5	0	5
Attention concentration	5.3±0.9	3	6
Language function	4.6±0.6	3	5
Orientation	5.8±0.6	4	6
K-BNT (total score)	78.5±22.3	15	99
K-MoCA (total score)	23.6±3.6	16	29

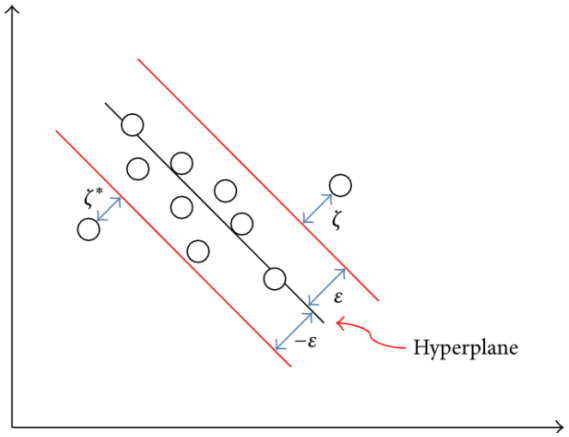


Fig. 3. The Feature Space of SVR [33].

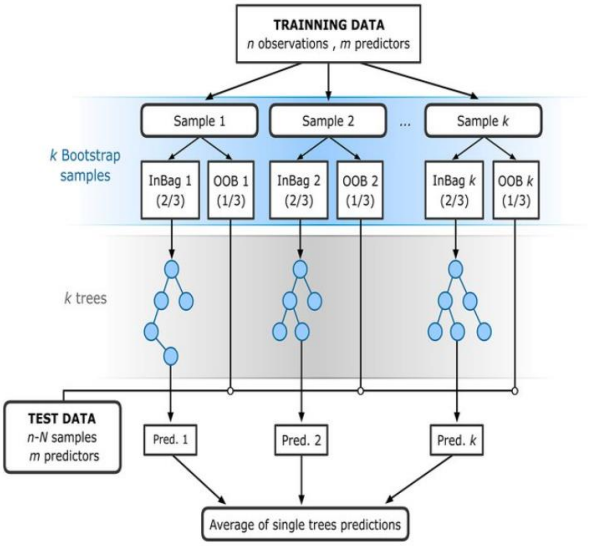


Fig. 4. Concept of a Random Forest [36].

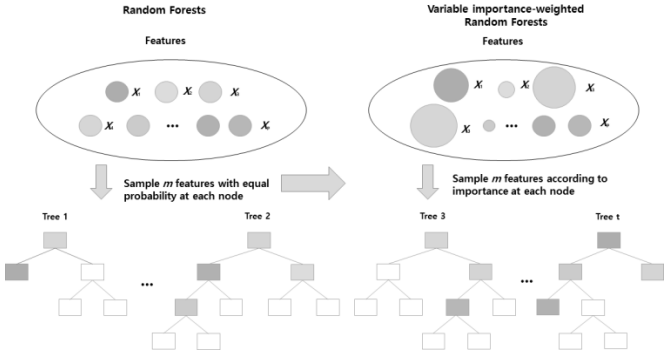


Fig. 5. Concept of a Weighted Random Forest [21].

III. Results

A. Development of Confrontation Naming Prediction Models for the Elderly in South Korea and Comparison of their Prediction Performance

Table II shows the root mean squared error of the confrontation naming prediction models for the elderly in South Korea, developed by using SVR, random forest, and weighted random forest. The results of this study defined a model with the lowest root mean squared error (%) as the model with the best prediction performance. As a result of the test for prediction performance, the random forest algorithm derived with 28.4% (a Root Mean Squared Error) was confirmed as the model with the best performance.

TABLE II. The Root mean Squared Error of the Confrontation Naming Prediction Models for the Elderly in South Korea

Model	Root Mean Squared Error (%)
SVR	31.1
Random forest	30.5
Weighted random forest	28.4

B. The Importance of Variables of the Final Model (Random Forest) for Predicting the Confrontation Naming of the Elderly Living in South Korea

Fig. 6 shows the importance of variables of the final model (random forest) for predicting the confrontation naming of the elderly living in South Korea. The final model confirmed that generative naming-meaning, generative naming-phonemes, memorizing numbers forward immediately, and memorizing numbers backward immediately, and memorizing numbers backward were the main variables with high weight for predicting the confrontation naming of the elderly. Among them, generative naming-meaning was the most important variable in the final model.

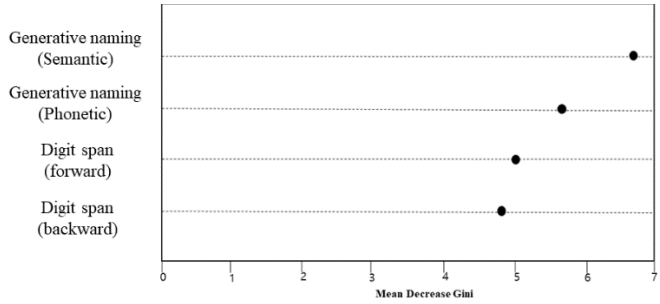


Fig. 6. Importance of Variables (Presenting the Importance of only the Top 4 Variables) of the Final Model (Random Forest) for Predicting Confrontation Naming of the Elderly Living in South Korea.

IV. Discussion

This study explored factors related to confrontation naming using SVR, random forest, and weighted random forest for the elderly in the community. The results of this study showed that the performance of confrontation naming was significantly associated with executive functions such as generative naming-meaning, generative naming-phonemes, memorizing numbers forward immediately, and memorizing numbers backward immediately. The results of this study agreed with the results of previous studies [38,39] based on the generalized precedence model (GLM), which showed that the performance of confrontation naming was significantly related to the generative naming the language domain of K-MMSE, number memorization (a test that measures working memory and attention), and the attention concentration domain. The results of this study implied that the healthy elderly without neurological diseases or dementia had a close relationship between the performance of confrontation naming and executive functions (e.g., generative naming and memorizing numbers immediately) [38] and executive functions could be major factors in predicting the performance of naming performance. In the future, longitudinal studies are needed to prove the causal relationship between cognitive functions and confrontation naming.

In this study, the CIU ratio of the picture description task was not a significant predictor of confrontation naming. Since the CIU analysis method is mainly used to analyze the language abilities of patients with central and peripheral nervous system damage such as aphasia and dementia, it could have a little impact on confrontation naming in this study, which targeted the healthy elderly in the community.

This study revealed that the root mean squared error of weighted random forests was the lowest when comparing the prediction performance using models based on SVR, random forest, and weighted random forest. Byeon et al. (2019) [21] developed a voucher service demand prediction model using weighted random forest, similar to this study, and showed that weighted random forest showed higher prediction accuracy than other machine learning methods. They suggested developing prediction models by using weighted random forest because weighted random forest giving more weight to good performing tree models showed better accuracy than the conventional random forest, which gives the same weight to all tree models.

V. Conclusion

This study, which analyzed the imbalanced data, also confirmed that weighted random forest has better predictive performance than random forest or SVR. It is believed that the weighted random forest will be more effective for developing prediction models for imbalanced y-variables. Future studies are needed to compare the prediction performance of weighted random forest with other machine learning models by calculating various performance indices such as sensitivity, specificity, and harmonic mean using data from various fields in order to prove the superior prediction performance of weighted random forest.

Acknowledgment

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

References

[1] Jeon, and S. Kwon, Health and long-term care systems for older people in the republic of Korea: policy challenges and lessons. Health Systems & Reform, vol. 3, no. 3, pp. 214-223, 2017.

[2] J. W. Han, T. H. Kim, . P. Kwak, K. Kim, B. J. Kim, S. G. Kim, J. L. Kim, T. H. Kim, S. W. Moon, J. Y. Park, J. H. Park, S. Byun, S. W. Suh, J. Y. Seo, Y. So, S. H. Ryu, J. C. Youn, K. H. Lee, D. Y. Lee, D. W. Lee, S. B. Lee, J. J. Lee, J. R. Lee, H. Jeong, H. G. Jeong, J. H. Jhoo, K. Han, J. W. Hong, and K. W. Kim, Overview of the Korean longitudinal study on cognitive aging and dementia. Psychiatry Investigation, vol. 15, no. 8, pp. 767-774, 2018.

[3] Y. S. Lee, S. D. Kim, H. J. Kang, S. W. Kim, I. S. Shin, J. S. Yoon, and J. M Kim, Associations of upper arm and thigh circumferences with dementia and depression in Korean elders. Psychiatry Investigation, vol. 14, no. 2, pp. 150-157, 2017.

[4] H. Byeon, Best early-onset Parkinson dementia predictor using ensemble learning among Parkinson's symptoms, rapid eye movement sleep disorder, and neuropsychological profile. World Journal of Psychiatry, vol. 10, no. 11, pp. 245-259, 2020.

[5] H. Byeon, Effects of grief focused intervention on the mental health of dementia caregivers: systematic review and meta-analysis. Iranian Journal of Public Health, vol. 49, no. 12, pp. 2275-2286, 2020.

[6] S. K. Kang, D. Y. Kim, D. I. Seok, H. J. Cho, and K. H. Choi, Studies on communication disorders in the elderly to improve their quality of life. Journal of Special Education & Rehabilition Science, vol. 40, no. 2, pp. 109-134, 2001.

[7] K. M. Yorkston, M. S. Bourgeois, and C. R. Baylor, Communication and aging. Physical Medicine and Rehabilitation Clinics, vol. 21, no. 2, pp. 309-319, 2010.

[8] J. Harwood, Understanding communication and aging: Developing knowledge and awareness. SAGE Publications Inc, New York, 2007.

[9] S. L. Danckert, and F. I. Craik, Does aging affect recall more than recognition memory?. Psychology and Aging, vol. 28, no. 4, pp. 902-909, 2013.

[10] L. Yang, and L. Hasher. Age differences in the automatic accessibility of emotional words from semantic memory. Cognition and Emotion, vol. 25, no. 1, pp. 3-9, 2011.

[11] A. K. Troyer, M. Moscovitch, and G. Winocur, Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology, vol. 11, no. 1, pp. 138-146, 1997.

[12] H. Byeon, Application of machine learning technique to distinguish Parkinson's disease dementia and Alzheimer's dementia: predictive power of Parkinson's disease-related non-motor symptoms and neuropsychological profile. Journal of Personalized Medicine, vol. 10, no. 2, pp. 31, 2020.

[13] M. A. Sager, B. P. Hermann, A. La Rue, and J. L. Woodard, Screening for dementia in community-based memory clinics. WMJ: official publication of the State Medical Society of Wisconsin, vol. 105, no. 7, pp. 25-29, 2006.

[14] G. Adler, S. Rottunda, and M. Dysken, The older driver with dementia: an updated literature review. Journal of Safety Research, vol. 36, no. 4, pp. 399-407, 2005.

[15] K. K. Tsoi, J. Y. Chan, H. W. Hirai, S. Y. Wong, and T. C. Kwok, Cognitive tests to detect dementia: a systematic review and meta-analysis. JAMA Internal Medicine, vol. 175, no. 9, pp. 1450-1458, 2015.

[16] J. S. Yu, A. Y. Xue, E. E. Redei, and N. Bagheri, A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder. Translational Psychiatry, vol. 6, no. 10, pp. e931-e931, 2016.

[17] H. Byeon, Developing a model for predicting the speech intelligibility of South Korean children with cochlear implantation using a random forest algorithm. International Journal of Advanced Computer Science and Applications, vol. 9, no. 11, pp. 88-93, 2018.

[18] H. Byeon, A prediction model for mild cognitive impairment using random forests. International Journal of Advanced Computer Science and Applications, vol. 6, no. 12, pp. 8-12, 2015.

[19] A. Iqbal, S. Aftab, Z. Nawaz, L. Sana, M. Ahmad, and A. Husen, Performance analysis of machine learning techniques on software defect prediction using NASA datasets. Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 300-308, 2019.

[20] L. Pourjafar, M. Sadeghzadeh, and M. Abdeyazdan, Combination of neural networks and fuzzy clustering algorithm to evalution training simulation-based training. Journal of Advanced Computer Science and Applications, vol. 7, no. 7, pp. 31-38, 2016.

[21] H. Byeon, S. Cha, and K. Lim, Exploring factors associated with voucher program for speech language therapy for the preschoolers of parents with communication disorder using weighted random forests. International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 12-17, 2019.

[22] D. Bruno, R. L. Koscik, J. L. Woodard, N. Pomara, and S. C. Johnson, The recency ratio as predictor of early MCI. International Psychogeriatrics, vol. 30, no. 12, pp. 1883-1888, 2018.

[23] S. Ellendt, B. Voβ, N. Kohn, L.Wagels, K. S Goerlich, E. Drexler, F. Schneider, and U. Habel, Predicting stability of mild cognitive impairment (MCI): findings of a community based sample. Current Alzheimer Research, vol. 14, no. 6, pp. 608-619, 2017.

[24] Kang, Y. Kim, H., & Na, D. L. Parallel Short Forms for the Korean-Boston Naming Test (K-BNT). Journal of the Korean Neurological Association, vol. 18, no. 2, pp. 144-150, 2000.

[25] Y. W. Kang, J. PARK, K. H. Yu, and B. C. Lee, A Reliability Validity, and Normative Study of the Korean-Montreal Cognitive Assessment (K-MoCA) as an Instrument for Screening of Vascular Cognitive Impairment (VCI). Korean Journal of Clinical Psychology, vol. 28, no. 2, pp. 549-562, 2009.

[26] Y. Kang, S. Jang, and D. L. Na, Seoul Neuropsychological Screening Battery (SNSB). Human Brain Research and Consulting Co, Seoul, 2003.

[27] H. H. Kim, and D. L. Na, Paradise Korean version-Western Aphasia Battery-Revised (PK-WAB-R). Paradise, Seoul, 2012.

[28] T. H. Yeom, Y. S. Park, K. J. Oh, J. K. Kim, and Y. H. Lee, Korean Wechsler adult intelligence scale (K-WAIS) manual. Handbook Guidence, Seoul, 1992.

[29] V. I. Sheikh, and V. A. Yesavage, Geriatric Depression Scale(GDS)：recent evidence and development of shorter version. In TL Brink(Ed), Clinical Gerontology：A guide to assessment and intervention. Haworth Press, New York, 1986.

[30] E. L. Lesher, J. S. Berryhill, Validation of the geriatric depression scale-short form among inpatients. Journal of Clinical Psychology, vol. 50, no. 2, pp. 256-260, 1994.

[31] M. B. Gerety, Jr. Williams, C. D. Mulrow, J. E. Cornell, A. A. Kadri, J. Rosenberg, L. K. Chiodo, and M. Long, Performance of case-finding tools for depression in nursing home： Influence of clinical and functional characteristics and selection of optimal threshold scores. Journal of the American Geriatrics Society, vol. 42, no. 10, pp. 1103-1109, 1994.

[32] A. J. Smola, and B. Schölkopf,. A tutorial on support vector regression. Statistics and Computing, vol, 14, no. 3, pp. 199-222, 2004.

[33] C. W. Chen, and Y. C. Chang, Support vector regression and genetic algorithm for HVAC optimal operation. Mathematical Problems in Engineering, vol. 2016, pp. 6212951, 2016.

[34] P. T. Noi, and M. Kappas, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors, vol. 18, no. 1, pp. 18, 2018.

[35] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics, vol. 8, no. 25, pp. 1-21, 2007.

[36] V. F. Rodríguez Galiano, M. Sánchez Castillo, J. Dash, P. Atkinson, and J. Ojeda Zújar, Modelling interannual variation in the spring and autumn land surface phenology of the European forest. Biogeosciences, vol. 13, pp. 3305-3317, 2016.

[37] E. J. Wagenmakers, and S. Farrell, AIC model selection using akaike weights. Psychonomic Bulletin & Review, vol. 11, no. 1, pp. 192-196, 2004.

[38] E. Higby, D. Cahana-Amitay, A. Vogel-Eyny, A. Spiro III, M. L. Albert, and L. K. Obler, The role of executive functions in object-and action-naming among older adults. Experimental Aging Research, vol. 45, no. 4, pp. 306-330, 2019.

[39] N. L. Saunders, and M. J. Summers, Longitudinal deficits to attention, executive, and working memory in subtypes of mild cognitive impairment. Neuropsychology, vol. 25, no. 2, pp. 237-248, 2011.

Towards the Development of Computational Thinking and Mathematical Logic through Scratch

Benjamín Maraza-Quispe[1], Ashtin Maurice Sotelo-Jump[2]
Olga Melina Alejandro-Oviedo[3], Lita Marianela Quispe-Flores[4]
Lenin Henry Cari-Mogrovejo[5], Walter Cornelio Fernandez-Gambarini[6], Luis Ernesto Cuadros-Paz[7]
Facultad de Ciencias de la Educación, Universidad Nacional de San Agustín de Arequipa, Arequipa-Perú[1, 3,5,6,7]
Facultad de Bioingeniería Universidad de Ingeniería y Tecnología Lima-Perú[2]
Facultad de Psicología, RR. II. y Cs. de la Comunicación, Universidad Nacional de San Agustín de Arequipa, Arequipa-Perú[4]

***Abstract*—Currently the need to provide quality education to future generations has led to the development of new teaching methodologies, within this fact the tools provided by information technologies have been positioned as the future of learning, in this sense, the learning to program is no longer considered a selective skill in the field of computing, being today a necessity for any student who wants to be competent in this globalized and dynamic world. Within this context, the present research aims to analyze to what extent the use of the Scratch programming language allows the development of computational thinking skills and mathematical logic. The methodology consisted of the application of programming fundamentals through Scratch 3.0 to an experimental group composed of 25 students who were randomly selected from a population of 100 students, the data collection was carried out through a test of logical reasoning standardized by Acevedo and Oliva and a test of levels of computational thinking standardized by González. According to the results, a significant difference is postulated in the performance of the students in both tests, having a more considerable improvement in the criteria: Loops, Control of Variables (CV), Probability (PB) and Combinatorial Operations (CB). Therefore, it is concluded by highlighting the importance of teaching basic concepts of Computer Science such as computational thinking and mathematical logic, since it contributes to the internalization of concepts when developing algorithms in problem-solving.**

***Keywords*—Scratch; computational thinking; logic reasoning; teaching**

I. Introduction

Nowadays, within the curricular networks of different Educational Institutions of Regular Basic Education (EBR), the implementation of IT tools in education is inappropriate due to a lack of competence of teachers in their use and also in part due to a lack of knowledge of the virtues that bring these strategies into the student's cognitive development, thus creating a perception of complexity towards programming for both students and teachers [1].

However, the virtue of the combination of conventional methodologies with IT tools are decisive, thus establishing young people who use computational thinking and mathematical-logical reasoning in their daily lives. According to [2] computational thinking can be defined as the ability to solve problems through capabilities such as algorithms and computational methods, this thinking is divided into four main processes: decomposition, abstraction, pattern recognition and algorithms. The aforementioned processes will have a fundamental role in the analysis of the benefits it brings to students, being these evaluated through the test designed by [3].

Similarly, the reasoning is made up of various capacities associated with mathematics, such as Pythagorean arithmetic or Euclidean geometry. On the other hand, beyond traditional conceptions, at present mathematical logic is considered as the reasoning that causes science through the validation of knowledge by the scientific method [4]. The importance of logical reasoning has always been valued in primary and secondary education. However, in the context presented, the following question arises: to what extent does the teaching of programming language through Scratch 3.0 develop Logical Reasoning and Computational Thinking in students of Regular Basic Education?

According to the existing bibliography, there seems to be a tentative answer that affirms said development of skills, but even more, there is an improvement in the acceptance of the error, since the success of this type of software lies in the development of the expected skills, thanks to the learning programming by goals [5].

II. Theoretical Framework

A. The Programming Languaje

In the world of computing, any technology or object that is called or has a computer or processor inside, works with a single language called binary code, that is, basically for a computer there are only ones and zeros.

Therefore, when it comes to the programming language, it is subdivided into two: high-level and low-level language; In this case, in the present work the first one will be taken as the central point, since the programs that will be exposed later use this type of programming language. These are closer to mathematical and natural language [4].

There are many types of programming languages below, in Table I which places Scratch within the types of programming languages:

TABLE I. TYPES OF PROGRAMMING LANGUAGE

Language type	Description	Software
C++	Intended for the development of programs or packages. Arduino Cross-platform Python programming language, ideal for beginners.	Eve online, Panda3D
Visual	It is a programming language in a growing state, mainly used to teach basic knowledge of the programming language.	Scratch, CODE
JavaScript	Designed to create programs that will be stored in web pages, it is also ideal for creating and implementing effects and actions.	Android Studio

B. Scratch 3.0 Software

Scratch defines itself as a program that allows you to create stories, games or animations; which can then be shared with the community, which provides positive feedback for users. In addition, it helps improve creativity and systematic thinking in young people [6].

In addition, among the numerous visual programming software, Scratch is the educational program par excellence, which provides a solution to the common abandonment of programming courses, caused by a perceived high difficulty that this activity entails [7].

Likewise, Scratch has a constructive and above all active teaching process, which generates in students a better experience when learning to program, and as Paper defended: programming languages must have a "low floor" and a "high ceiling" In other words, it should not be a challenge to understand how to start programming, however the possibilities must be gigantic [7].

C. Computational thinking

In this modern and changing world, technology is an important part of the development of new methods to improve the productivity of certain products [8].

Computational thinking is that way of solving problems, using computational methods or methods normally used by technology such as algorithms [2].

This process consists of several parts to follow, which are developed through a specific reasoning called computational logic.

Computational thinking is of great importance for current demands, these being characterized by the constant problems that employees are subjected to, for example. In short, computational thinking is imperative if you want to "survive" in this globalized world. Because the teaching of this system provides the population with a tool so that, as mentioned above, it improves production and efficiency in any type of work environment and in daily life.

In this case, greater importance will be given to decomposition as a fundamental part of computational logic and therefore of computational thinking.

D. Logical Reasoning

Logical reasoning is composed of various capacities associated with mathematics, such as Pythagorean arithmetic or Euclidean geometry. On the other hand, beyond traditional conceptions, at present mathematical logic is considered as the causal reasoning of science through the validation of knowledge by the scientific method [9].

It is extremely important, the importance of logical reasoning has always been valued in primary and secondary education. However, in recent years, the development of these mathematical skills is the same, the methods can be varied, in this case we will study the impact of IT on the learning of mathematical logic.

E. Teaching Scratch 3.0 to EBR Students

Reference [7], dealing specifically with Scratch, as this is a type of block-oriented visual programming language, it has a constructive teaching process, ideal to initiate students to programming. This gradual process accompanied by the development of expected skills through a programming system for individual goals makes Scratch 3.0 the programming software par excellence for young people.

On the other hand, it is worth emphasizing what the literature in recent years says about the development of both thoughts and their relationship with Scratch: first, logical reasoning is composed of various capacities associated with mathematics, such as Pythagorean arithmetic or Euclidean geometry. On the other hand, at present, mathematical logic is considered as the reasoning that gives rise to science through data validation [4] that is why authors such as [10] present a curriculum for students that are oriented to their true needs; therefore, globalization is considered as the first axis.

As for how Scratch manages to develop the aforementioned capacities, everything lies in the principles with which it was devised, below in Table II, the three fundamental principles are presented according to [7].

TABLE II. SCRATCH 3.0 PRINCIPLES [7]

Scratch Principles	Description
The programming language must be playful	The ease with which you can try different options to complete a certain action is essential to improve the experience of what it is to program.
The programming language must be meaningful	When a person wants to learn something new, one of the best ways to do it is if the activity is meaningful to the person, that is, it has a certain degree of relevance and authenticity for each user. Scratch is designed precisely to meet this requirement; it is diverse and personal at the same time.
The programming language must be social	Scratch is closely linked to its website, since in this way a good community has been consolidated around the MIT platform so that each user, regardless of age, can share your work with the whole world with just one click; being able to receive the necessary improvements for their animation, comic strip, game or project in general. In short, thanks to the Scratch website, you have personal feedback for each user.

The model proposed by [7] is achieved thanks to the Scratch interface. The same one that had an original design motivated to meet the learning needs and, besides, create interest in children and young people. Thus, in the first instance, Scratch began to be used in places outside the classroom, although it would inevitably reach the curricula of thousands of schools around the world, due to this invitation to exploration and exchange with peers [11].

F. The Literature on Scratch and its Impact on Teaching

Especially in the last decade, studies have been conducted at schools about student acquisition in programming and computational thinking skills. Scratch is regarded as a useful tool in teaching programming or ensuring that students acquire computational thinking skills [12-13].

Scores obtained from assessing Scratch projects via Scratch web tool and students' Computational Thinking Levels do not differ based on gender. In other words, gender is not influential on students' project assessment scores. While literature presents extensive proof for the impact of gender on student's characteristics related to computers or programming [14-15]. There is a significantly high relationship between students' Scratch skills and their computational thinking skills. In other words, development in students' programming skills in Scratch will cause similar increases in their computational thinking skills or improvements in their Computational Thinking Levels will generate increases in their Scratch skills. Literature provides extensive proof that the process of programming is not mechanical, but a thinking discipline [16].

According to [17], in an investigation developed which aimed to compare the scores of fifth-grade students obtained from Scratch projects, the scores obtained of the Scale of Levels of Computational Thinking and Examine this comparison in terms of different variables. A correlational research model was used in the study in which 31 students participated. Students were taught basic programming using Scratch for 6 weeks' period. At the end of the training, the students' scheduling skills were measured through the Dr Scratch web tool. Computational thinking skills were measured using Scale of levels of computational thinking that includes 5 factors: creativity, problem-solving, algorithmic thinking, collaboration and critical thinking. The data were analyzed for internal reliability to calculate the reliability of the scale. Cronbach's alpha reliability coefficient was found to be 0.809. It was found that the scores obtained by the students by use of any of the measurement tools did not differ depending on gender or period of computer use, however, a significant high-level relationship was observed between students' programming skills with Scratch and your computational thinking skills.

III. Methodology

La metodología utilizada está enfocada a un estudio cuantitativo para determinar puntos concluyentes sustentados en datos numéricos sometidos a tratamiento estadístico para corroborar su validez, de esta manera es que se opta por este tipo de metodología en lugar de un estudio cualitativo, con el fin de buscar la mayor certeza de que el tratamiento es el principal motivo de los resultados y no factores externos, además de ser más relevante al momento de replicarlo en diferentes contextos mundiales.

A. Objective

Analyze the development of computational thinking and mathematical logic through the visual programming language Scratch.

B. Population and Sample

The total population is made up of 100 students in the third grade of Regular Basic Education, of which 25 students were selected for the experimental group through a simple random sampling. With the sample described above, the constant evaluations to which they will be subjected will be important, through tests designed to quantify the use of computational thinking in solving problems.

C. Process

Learning sessions were designed consisting of a minimum of four lessons, each lasting sixty minutes and divided into four parts: purpose, where the goal of the class is expressed; development, where we proceed with the explanation of the blocks and the structure of visual programming; evaluation, a moment in which an exercise with the name of "Challenge" is proposed and finally exit, the final section where all the concepts are recovered and conclusions are drawn.

Finally, the sample was evaluated with a pre-test, before applying the programming lessons and at the end of the lessons, it was completed with a post-test; both evaluations as a data collection instrument.

D. Data Collection Instrument

Two collection instruments were used due to the two fields of study: mathematical logic and computational thinking. In the case of the first, what was proposed by Tobin and Copie (1981), and their study "Test of Logical Thinking" (TOLT), and the subsequent conversion and validation to the Spanish language, carried out by [18], the same that assesses through open and closed questions five criteria on logical reasoning: proportionality (PP), control of variables (CV), probability (PB), correlation (CR) and combinatorial operations (CB).

In the case of computational thinking, it was based on Gonzales' study, in which, through a test of 28 multiple-choice items, it is aimed at the standard quantification of the levels of computational thinking in the subjects, in solving problems by helping each other. With computational concepts: Basic Directions, Loops (repeat times), Loops (repeat until), Simple Conditional, Compound Conditional, While (while) and Simple Functions.

The test lasts 45 minutes and as an objective population to students from twelve years to fifteen years [3].

IV. Analysis and Results

A. Data Collection, Classification and Analysis

According to the results obtained, the performance of the sample between criteria was considerably varied, so it was decided to carry out a detailed study by the criterion of each collection instrument: in the case of mathematical logical reasoning, the marks were classified into five criteria that are

those presented by the author of the test and in the case of computational thinking, the same was done but in seven evaluation criteria. In this way, we seek to obtain more accurate and specific data on in which fields there is a significant improvement after treatment (Scratch 3.0 teaching).

The final classification table for the Logical Reasoning Test (TRL) is shown below. With a maximum score of 50 per criterion.

From the data processed in Table III, not all the criteria evaluated by the Acevedo and Oliva test had the same change. The second and third criteria are specifically rescued, which explain: use of variables and proportionality where there is a more significant change concerning the others. This is due to the same process of using Scratch [7].

Regarding the second field of study, as mentioned above, the same principle applies with the difference that each criterion of the Computational Thinking Test has a maximum score of 100.

As analysed from Table IV, specifically, in the second and third criteria, there is a more significant improvement. These criteria measure the ability of students to use loops both in numerical repetitions (repeat how many) and in repetitions with conditionals at the end (repeat until). This concept is related to what [19]: (…) When writing code, students learn how to organize a process, recognize routines or repetitions and discover errors in their computational thinking when their program does not work according to the idea or expectation with which it was conceived. All of them are key features of computational thinking. (…). Which would explain the best performance in the criteria most closely related to the processes that define computational logic and problem solving, fundamental pillars in computational thinking.

TABLE III. EVALUATION CRITERIA IN THE LOGICAL REASONING TEST

Criteria	Total of the results obtained	
	Pre-test	Post-test
Proportionality	48	50
Variables Control	26	37
Probability	27	36
Correlationt	27	29
Combinatorial Operations	33	33

TABLE IV. EVALUATION CRITERIA IN THE COMPUTATIONAL THINKING TEST

Criteria	Total of the results obtained	
	Pre-test	Post-test
Basic Directions	99	100
Loops (repeat times)	41	71
Loops (repeat until)	39	69
Simple conditional	42	50
Compound Conditional	33	48
While	40	55
Simple Functions	32	42

B. Validation of the Proposal

To validate the presented proposal, the use of analysis of variance for paired samples was used, through this statistical treaty it is sought to conclude whether the treatment carried out had a significant impact or otherwise the results are not good enough to affirm that teaching Scratch improves logical reasoning and computational thinking.

According to Fig. 1, the results show a greater significant effect between the pre and post-test.

According to Fig. 2, the results show a greater significant effect between the pre and post-test, based on the postulate of [18] where there is evidence of a greater development by criteria in the post-test.

On the other hand, although there is a considerable improvement in the criteria explained, analyzing the means of both groups of data, the difference of is 1.16. Using the studied authors and analyzing Fig. 1, it can be explained: that a possible cause may be the orientation of the methodology in the treatment, due to the educator's determining degree in logical reasoning, as concluded by [20]: Development of Logical Reasoning: (…) The teacher must provide his students with the necessary tools to learn, thus mediating their learning. It is an urgent need to promote the development of capacities and values in the classroom. (…)

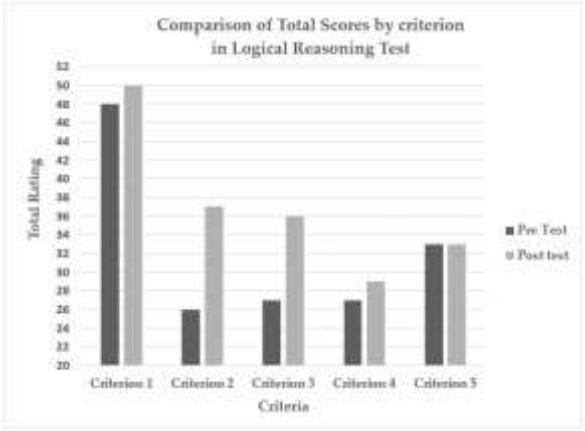


Fig. 1. Comparison per the Criterion of Results in the Logical Reasoning Test.

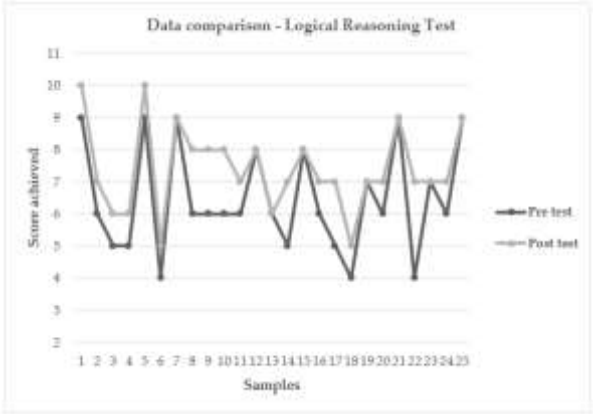


Fig. 2. Comparison of General Results in the Logical Reasoning Test.

Likewise, developing in the case of the first and fifth criteria, when dealing with Propositions and Combinatorial Operations, the characteristics of the sample and the treatment time may have influenced the results, because they are more complex concepts than when comparing them with age. The sample mean, is not yet fully internalized.

A considerable difference between the pre and post-test is observed in Fig. 3 and Fig. 4, which promises good results in the treatment. However, the improvement has not been uniformed in all the criteria, see Table III, in which criteria such as basic directions and simple functions have fairly close results.

In conclusion, the treatment in the selected sample has had the expected impact, significantly increasing performance within the TPC, based on [3] study, thus validating the contribution of various authors in the development of Computational Thinking through the teaching of Scratch.

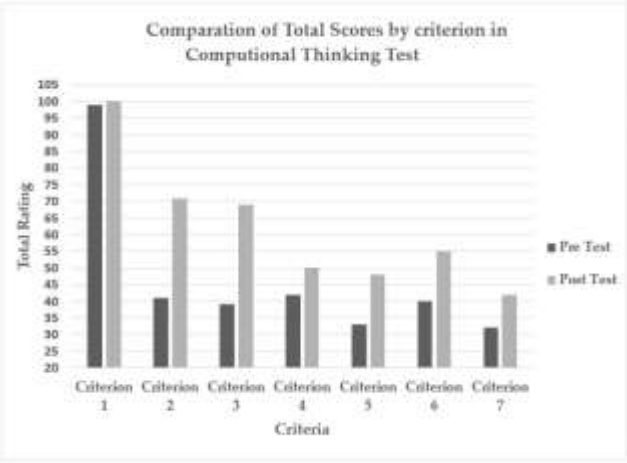


Fig. 3. Comparison per Criterion of Results in the Computational Thinking Test.

Fig. 4. Comparison of General Results in the Computational Thinking Test.

C. T-student for Two Related Samples

The statistical treatise used consists of the formulation of two hypotheses: the null hypothesis and the alternative, depending on the results of the treaty, one of the hypotheses is approved or refuted.

1) Analysis by criterion within the logical reasoning test:

Ho = There is no significant difference in the means of the results in the TRL before and after the treatment.

Ha = There is a significant difference in the means of the results in the TRL before and after the treatment.

$\alpha = 5\%$

According to the data presented in Table V, given that the resulting P value is less than 5% or 0.05, then the null hypothesis is rejected and the alternative hypothesis is accepted, in other words, if there is a significant difference after treatment.

2) Analysis by criterion within the logical reasoning test: Once the data were validated within the TRL, the significance was then studied by criterion. In this way, more specific conclusions were obtained in singular fields and the impact of the treatment (teaching of Scratch 3.0 in third-year high school students) in the sample.

According to the data presented in Table VI we have

TABLE V. T-TEST FOR TWO PAIRED SAMPLES IN THE LOGICAL REASONING TEST

	Variable 1	**Variable 2**
Average	6.44	7.4
Variance	2.76	1.45
Observations	25	25
Pearson's correlation coefficient	0.86	7.4
Hypothetical difference of means	0	
Degrees of freedom	24	
T statistic	5.70	
P (T <= t) one tail	3.49	
Critical value of t (one-tailed)	1.71	
P (T <= t) two tails	**0.0000070**	
Critical value of t (two-tailed)	2.063	

TABLE VI. T-TEST BY CRITERION IN THE LOGICAL REASONING TEST

Criteria	**P-Value**	**alpha**	**Significant difference**
Proportionality	48	50	NO
Variables Control	26	37	SI
Probability	27	36	SI
Correlationt	27	29	NO
Combinatorial Operations	33	33	SI

3) General test of the computational thinking test:

Ho = There is no significant difference in the means of the results in the TPC before and after the treatment.

Ha = There is a significant difference in the means of the results in the TPC before and after the treatment.

$\alpha = 5\%$

Table VII shows that, as in the case of the logical reasoning test, the value of P is less than alpha, which is why it is stated that there is a significant difference.

As shown in Table VIII, there is a significant difference in the means of the Computational Thinking Test results. Therefore, it is concluded that the teaching of Scratch does have significant effects on the development of Computational Thinking.

These indicators of learning behaviors in virtual learning environments are very important for self-regulation and reflection of students and teachers within their teaching and learning context. Likewise, teachers could provide very effective feedback by knowing the indicators of learning behavior in which they have weaknesses. That is why teachers considered that these data could help in the redesign of their courses [21].

TABLE VII. T-TEST FOR TWO PAIRED SAMPLES IN THE COMPUTATIONAL THINKING TEST

	Variable 1	**Variable 2**
Average	13.04	17.4
Variance	19.12	22.42
Observations	25	25
Pearson's correlation coefficient	0.28	-
Hypothetical difference of means	0	
Degrees of freedom	24	
T statistic	-3.98	
P (T <= t) one tail	0.00	
Critical value of t (one-tailed)	1.71	
P (T <= t) two tails	0.000553121	
Critical value of t (two-tailed)	2.06	

TABLE VIII. T-TEST BY CRITERION IN THE COMPUTATIONAL THINKING TEST

Criteria	**P-Value**	**alpha**	**Significant difference**
Basic Directions	0.16	0.05	NO
Loops (repeat times)	0.00	0.05	SI
Loops (repeat until)	0.00	0.05	SI
Simple conditional	0.31	0.05	NO
Compound Conditional	0.07	0.05	NO
While	0.08	0.05	NO
Simple Functions	0.06	0.05	NO

V. DISCUSSION AND CONCLUSIONS

The present investigation concluded that the treatment used in the sample has managed to have a significant impact on the results of both fields of study: computational thinking and logical reasoning.

In the first place, in the case of Computational Thinking, the benefits of programming teaching in students, with abilities related to problem-solving, as well as of all ages, are affirmed, without making any type of distinction because the activity it adapts to the individual capacities of each person. Leaving aside the development of the capacities evaluated during the present work, the improvements also focus on soft skills within the students, bringing them closer to a modern and dynamic environment, typical of the world of work.

Secondly, the learning of mathematical logic has been confirmed by statistical analysis, which affirms the advantages of the implementation of IT in apparently foreign areas such as school mathematics, all this achieved through the application of the TRL a standardized test. In this way, the teaching of Scratch is recommended more specifically to improve the performance of students in the subject of control of variables and probability.

Third, the teaching environment that Scratch creates is worth emphasizing, especially with a view to the future implementation of virtual environments in thousands of schools around the world. This software allows students to work as a team in addition to allowing a sociocultural exchange that enriches their perspective of the world.

In this way, to a large extent, the teaching of Scratch 3.0 has allowed the development of computational thinking and logical reasoning in high school students, which augurs the good relationship between IT and Education, innovating and giving away new learning tools by which, the new generations are formed with modern and fundamental capacities in today's world.

VI. RECOMMENDATIONS AND FUTURE WORK

To obtain a higher percentage of reliability in the results, it is recommended to obtain the data through a qualitative approach to observe qualities and behaviors and work with students from a more personalized perspective using some Artificial Intelligence techniques [22].

REFERENCES

[1] Maraza Quispe, B. (2011). Influencia de un entorno multimedia de simulación por computadora en el aprendizaje por investigación de la Física. Nuevas Ideas en Informática Educativa, TISE.

[2] Moreno León, J. (23 de marzo de 2014). Programamos. Obtenido de Videojuegos y "apps": https://programamos.es/que-es-el-pensamiento-computacional/.

[3] Román Gonzalez, M., Pérez Gonzalez, J. C., & Jiménez Fernández, C. (2015). Test de Pensamiento Computacional: diseño y psicometría general. Congreso Internacional sobre Aprendizaje, Innovación y Competitividad.

[4] Hernandez, L. (2004). Fundamentos de la programación. Madrid, España.

[5] Sánchez Rey, A. (2016). UNIVERSIDAD DEL PAÍS VASCO. Obtenido de Archivo Digital Docencia Investigación: https://addi.ehu.es/bitstream/handle/10810/20672/TFG%20Ane%20Sanchez.pdf;jsessionid=B9E714318BADDCB1081E77929F588C86?sequence=2.

[6] Scratch-MIT. (s.f.). Scratch. Obtenido de Acerca de Scratch: https://scratch.mit.edu/about.

[7] López-Escribano, C., & Sánchez-Montoya, R. (2012). Revistas UM. Obtenido de https://revistas.um.es/red/article/view/233521/179471.

[8] Colegio de Bachilleres. (2008). Lógica Computacional y Programación. Mexico.

[9] Sangüillo Fernandez-Vega, J. (2008). El pensamiento lógico matemático. Madrid: Ediciones Akal, S.

[10] Chamorro, C. (2011). Revistas UM. Obtenido de https://revistas.um.es/educatio/article/view/132961/122661.

[11] Maloney, J., Resnick, M., Rusk, N., Silverman, R., & Eastmond, E. (Noviembre de 2010). The Scratch Programming Language. ACM Trans, 15.

[12] S. Catlak, M. Tekdal. and F. Baz. The Status of Teaching Programming with Scratch: A Document Review Work. Journal of Instructional Technologies & Teacher Education, 2015, 4(3), 13-25.

[13] O. Ozyurt. and H. Ozyurt. A Study for Determining Computer Programming Students' Attitudes Towards Programming and Their Programming Self - Efficacy. Journal of Theory and Practice in Education, 2015, 11(1), 51-67.

[14] S. Toker. An Assessment of Pre-Service Teacher Education Program in Relation to Technology Training for Future Practice: A Case of Primary School Teacher Education Program, Burdur.

[15] (Unpublished master's thesis), 2014, Gradute School of Education, METU.

[16] A.P. Ambrosio. F. M. Costa. L. Almeida. A. Franco. and J. Macedo. Identifying cognitive abilities to improve CS1 outcome. Paper presented at the Frontiers in Education Conference (FIE), 2016.

[17] Ali OLUK, Özgen KORKMAZ,"Comparing Students' Scratch Skills with Their Computational Thinking Skills in Terms of Different Variables", International Journal of Modern Education and Computer Science(IJMECS), Vol.8, No.11, pp.1-7, 2016.DOI: 10.5815/ijmecs. 2016.11.01.

[18] Acevedo, J. A., & Oliva, J. (1995). Validación y aplicaciones de un test de razonamiento lógico. Revista de psicología general y aplicada, 339-351.

[19] Valverde Berrocoso, J., Fernández Sánchez, M. R., & Garrido Arroyo, M. (2015). El pensamiento computacional y las nuevas ecologías del aprendizaje. Revista de Educación a Distancia(46), 18.

[20] Cerillo MM. (2002). Enseñar a Pensar: Desarrollo de Razonamiento Lógico. Aulo Abierta.

[21] Maraza-Quispe, B., Alejandro-Oviedo, O., Choquehuanca-Quispe, W. (2020). Towards a Standardization of Learning Behavior Indicators in Virtual Environments. International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11. From https://thesai.org/Downloads/Volume11No11/Paper_19-Towards_a_Standardization_of_Learning_Behavior.pdf.

[22] Maraza, B. (2016). hacia un Aprendizaje Personalizado en Ambientes Virtuales. Campus Virtuales, Vol. 5, num. 1, pp. 20-29. en www.revistacampusvirtuales.es.

Survey of Centralized and Decentralized Access Control Models in Cloud Computing

Suzan Almutairi[1], Nusaybah Alghanmi[2], Muhammad Mostafa Monowar[3]
Technical and Vocational Corporation, Saudi Arabia[1]
Department of Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[2, 3]

***Abstract*—In recent years, cloud computing has become a popular option for a number of different businesses sectors. It is a paradigm employed to deliver a range of computing services, such as sharing resources via the Internet. Security issues in cloud computing necessitates the need for a mechanism to keep the system safe and reliable. An access control mechanism is one that permits or denies access to cloud services. This paper presents a survey of access control models in Cloud Computing. Several existing surveys on access control mechanisms in cloud computing mainly focused on traditional access control models and encryption-based access control models while the others focused on applying blockchain technology in cloud access control. However, access models possess different characteristics, such as the system's reliance on a centralized cloud trusted system administrator to manage the access policy or adopting decentralized approach. This paper reviews and analyses existing access control mechanisms in cloud computing, based on centralized and decentralized access control models, provides detailed comparisons on each model's advantages and limitations, and discusses the challenges of, and future research direction for access control.**

Keywords*—Cloud computing; access control; cloud security; centralized; decentralized

I. Introduction

Cloud computing has recently become the information technology (IT) foundation for many companies and organizations due to many its benefits, such as its interoperability, mobility, and cost effectiveness. Cloud computing technology refers to the virtualization of the IT infrastructure, including the hardware, software, and networking. This IT infrastructure is associated and designed together to provide the cloud services to the end user via the Internet [1-3]. Cloud computing includes three service models:

- Software as a service (SAAS): This model offers a variety of applications to the user;
- Platform as a service (PAAS): The model offers infrastructure as an environment for the developer to develop their own services or applications;
- Infrastructure as a service (IAAS): The model offers a virtualized resource, such as database services, a virtual machine, and a storage service to the user.

Securing the cloud environment is a critical issue, due to the unique characteristics of cloud computing, such as multitenancy and elasticity of the sharing of resources. It requires an access mechanism to ensure confidentiality, integrity, and availability for cloud data [4, 5]. Access control is considered to be the first line of defense of the system, and allows authorized users to gain access to the protected information and system resources, as well as denying unauthorized users access to the same.

In cloud computing, access control permits cloud users to access specific applications, or to protect data privacy, and can also be applied to protect the cloud user's resources. Finally, it gives the correct access permissions to each level of service. For example, in IAAS, it separates access permission to the guest's virtual machine from the host operating system. Access control in the cloud can be formed as either centralized or decentralized access control models. The former depends on a central authority to manage the access policies and key generation, while the latter depends on a multi-authority to manage the keys, and to store encrypted resource and access policies.

A. Contribution

Several recent works presented survey papers concerning the traditional access control models and encryption-based access control models for cloud computing, [1, 6]. Also, Xie et al. [7] focused on blockchain technology in the cloud, and its associated issues. However, no extant work presented a survey from the way the access control policies are managed (centralized and decentralized). The contributions of this paper are therefore as follows:

- We provide a taxonomy, based on centralized and decentralized access control models;
- We present detailed comparisons of the existing solutions of centralized and decentralized access control models, and the strengths and the limitations of each;
- We discuss the challenging issues with the existing access control models, and provide future directions.

B. Motivation

The proposed access control models usually possess two main characteristics [8]: first, they require one or more centralized centers to store or manage different data, such as user identities, cryptographic keys, and access rights etc. Second, all three cloud service models require a cloud trusted system administrator to manage the access rights and authorization process for other users. Consequently, there are two issues: first, an attack on the centralized center, causing single point of failure resulting in data compromise. In this case, the attacker may tamper with the data access, steal the

resources, or cause other forms of damage. Second, a malicious cloud security administrator could use their authority to access resources illegally, or to tamper with legal users' access rights, which would engender a loss of confidence and trust in the cloud.

In order to reduce the effects of these issues, the researchers start to look at new techniques to decentralize the cloud storage of access control. The blockchain or multiple distributed authority are an example of these techniques. The blockchain technique is more difficult to manage compared to centralized access control. Despite that, they are more secure in key distribution, file information, etc. Another issue is blockchain is slower than centralized access control because of the blockchain design nature [9].

This study therefore reviews and analyses the relevant literature on existing access control mechanisms in cloud computing that concern centralized and decentralized access control models, and assesses their advantages and disadvantages.

The rest of this paper is organized as follows: Section II introduces the background to access control models, blockchain technology, and smart contract techniques. Section III presents the existing solutions for access control in centralized and decentralized models in cloud computing, while Section IV discusses the challenges of, and potential future research direction for access control models in cloud computing. Finally, Section V concludes the paper.

II. Background

This section provides the background to this paper, including the concept of access control, blockchain technology, and smart contracts.

A. Access Control

This subsection presents the basic elements of access control, access control models, and encryption based access control.

1) Basic elements: There are three elements involved in the access control model, namely subject, object, and access rights [10]. The subject is the entity (users or applications) that can access an object, while the object is a resource, such as files or directories, that requires access. Lastly, access rights include the access policy, such as read or write, from the subject to the object.

2) Traditional access control models: Access control models are described as either discretionary or non-discretionary, and there are three main types, namely discretionary access control (DAC), role-based access control (RBAC), and mandatory access control (MAC) [11]. In DAC, the object's owner is required to specify the subject and the associated privileges and access policy. In RBAC, the access policy is based on the role of the subject. In MAC, the subject sensitivity label is compared with the object sensitivity label, and the former must be equal to or higher than the latter.

3) Encryption based access control: Cryptography algorithms are used to store and protect data as ciphertext, in order that the data remains secure in the cloud. Thus, encryption-based access control assists with achieving complementarity by combining cryptography algorithms with policy-based access control [6]. Attribute-based encryption (ABE) is a leading model of encryption-based access control.

In ABE, the identity of the user is defined by a set of attributes, and is categorized into key-policy ABE (KP-ABE) and ciphertext-policy ABE (CP-ABE) [6]. In a CP-ABE system, a set of attributes are assigned to the user's private key and access structures are bounded in ciphertext. For the decryption of the ciphertext, the user's attributes are matched with the access structure. Meanwhile, in KP-ABE, in the phase of generating the key, the access structure is associated with the user's private key.

B. Blockchain Technology and the Smart Contracts Technique

This subsection presents the basic characteristics of blockchain technology and the smart contracts technique. A blockchain is a sequence of connected blocks. In basic terms, the block is a data container that holds multiple fields within it, such as data transaction, the hash value of the current block, the hash value of the previous block, and the timestamp [12]. It is like a database that holds every transaction as a record. Furthermore, the blockchain has multiple characteristics [7] such as:

- Decentralized: The blockchain is a distributed network, which means that everyone who has been authenticated in the blockchain can participate, and can maintain the entire blockchain;
- Immutable: It is theoretically impossible to change or edit a transaction in the blockchain. This is because of the consensus mechanism, in which every block needs to compute a cryptography puzzle within 10 minutes to be added as a new chain. This new block is broadcast to the network, and other participants verify the correctness of the new block and the transactions it contains. After the block has been verified, it is added to the chain. In addition, a new hash value is generated for the current block, and a hash value for the previous record. The operations are on-going. As a result, any tampering with any block hash value will have to change all the subsequent blocks within a specific timeframe. Therefore, any tampering in any block can be detected immediately;
- Anonymity of user identity: The block in the chain has a unique wallet address. These addresses are generated using public and private pair keys. Every transaction occurs using the user wallet address.

In the same context, the smart contracts technique is a type of computer program that can be auto-generated, based on author codes, and cannot be modified once is generated and deployed, without the need for human interaction. It is used in the blockchain to store the encrypted contract that contains the keyword index (hash value), and other related data. The encrypted keyword, which is a unique hash identifier, helps in searching the service quickly and in retrieving only the correct

results. Consequently, the fee of the service is calculated and deducted from the customer contract. This technique solves the problem in centralized cloud access control of the retrieval of incorrect results, or no results at all, to save cloud computational resource costs [13].

III. Access Control Models in Cloud Computing

This section presents the current centralized and decentralized access control solutions in cloud computing. The former require a central authority to manage the policy and the keys, and to store the data, while the latter require a distributed authority.

A. Centralized Access Control

In the centralized access control model, there are four entities: the data owner (DO), central authority (CA), cloud server (CS), and data user (DU), as shown in Fig. 1. The CA is a trusted center, responsible for the centralized concept, which manages the access policy and generates the keys for the DU and the DO. The DO uploads the relevant resources and stores the data to the CS, while the CS stores the data and provides a transmission service between the DU and the DO. Finally, the DU receives and downloads the required data, according to the access policy.

Liu et al. [14] suggest an online/offline CP-ABE scheme in mobile cloud computing for improving the computational overheads of E-Healthcare Records (EHR). Offline encryption allows major computation, and ensures that the computation online encryption task is reduced. In the scheme proposed, the data encryption is formed of two phases, namely online and offline encryption. In the offline encryption, the EHR owner (Internet of Things (IoT) device) encrypts the data so that intermediate ciphertext is produced, which is subsequently used in the online encryption. In the online encryption, once the data is ready, the access policy is specified by the owner, and the final ciphertext is sent to the cloud. The scheme proposed is superior to other schemes in regard to the online encryption and decryption costs. However, the trade-off between the computation time and storage space should be addressed.

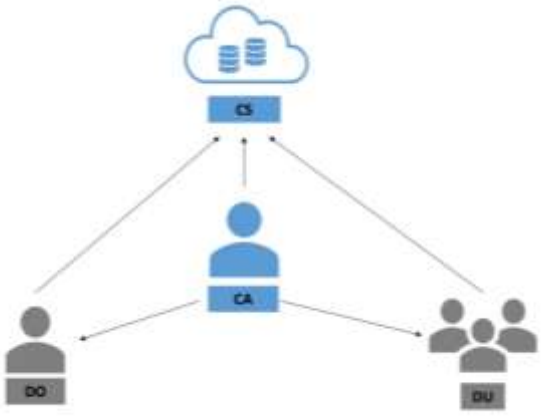


Fig. 1. Centralized Access Control Model

Meanwhile, Lin et al. [15] propose the use of PriGuarder to protect data privacy in the cloud. The approach proposed has three stages: DU registration, data creation, and DU access. In each stage, there are two access modes: direct, or anonymous. In the direct access mode, the operation occurs between the DU and the CS. In the anonymous access mode, the operation occurs between the DU and a trusted third party (TTP), with the TTP converting the DU data, identity, or access policy using an attribute fuzzy grouping (AFG), and sends the converted result to the CS. In the DU registration, the DU chooses the access mode to generate the DU identity. Then, in the data creation stage, the DO transmits their data, along with a statement of policy rights, and the chosen access mode. Finally, the verification of the DU occurs in the data access stage, according to the access mode selected. The key power in PriGuarder is the AFG method, which protected user privacy. However, the study fails to consider that a possible malicious TTP might contravene the users' privacy.

Generally, the hierarchy CP-ABE classifies the related attributes into different attribute trees. In their study, Li et al. [16] provide an efficient extended file hierarchy CP-ABE scheme (EFH-CP-ABE), in which they overcome the issue of encrypting multiple files with a similar access level. As shown in Fig. 2 [16], the EFH-CP-ABE scheme has four entities, namely the DO, the CA, the CS, and the DU. The CA generates three keys: a secret master key, a public key, and a private key for each user. The CS provides the transmission service and stores the ciphertext in the storage, while the DO stores and shares the data with the CS. Finally, the DO entity divides the data (m) into different blocks, such as m_1, m_2, and mi, and provides different session keys (sk) such as sk_1, sk_2, and sk_i. The DO randomly encrypts (E) the block mi with ski to produce $E_{sk}(M)= \{E_{sk1}(m_1), E_{sk2}(m_2), E_{sk3}(m_3)\}$, which is then stored in the CS. Finally, the DU downloads the ciphertext and decrypts part or all of the ciphertext, according to the node level's particular attributes. The results of the study demonstrate that the access control scheme achieved security and flexibility for cloud storage users. However, the computation time required for the encryption and decryption in the scheme requires improvement.

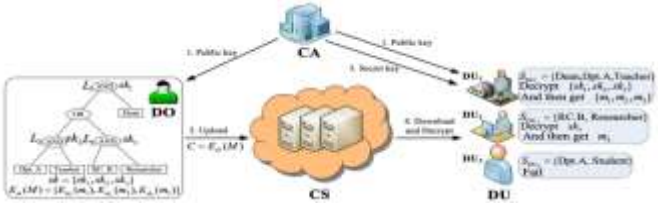


Fig. 2. EFH-CP-ABE Architecture [16].

Meanwhile, Jamal et al. [17] suggest a solution that provides a backup authority node in case of failure, and an efficient method of data access. This is an agent-based ABE access control method that employs the encrypted policy ABE mechanism, and has seven entities. First, the certification authority functions as a certificate issuing agent. Second, there are multiple attribute authorities, each with responsibility for delivering the encryption and decryption keys to the DO and the certified DUs. Third, the client site agent passes the user requested data to the server site agent to retrieve their data

from the cloud storage. Fourth, the server site agent requests are processed at the server-site. Fifth, the client storage server provides storage services and computational resources to the DU and DO. Sixth, the authorized agent keeps track of the neighbour nodes; should there be a certificate authority failure, the authority's back node is activated, subject to a request. Finally, the request handler has responsibility for the data access processing, using shared cache memory scheduling. The approach proposed is secure against malware injection, identity theft, collusion, and certification authority failure attacks. Additionally, the reading response times when accessing the cloud data are improved. However, the process of selecting the appropriate backup authority node needed to be secure.

In their work, Anilkumar and Subramanian [18] propose an algorithm called PB-FGAC, which combines a predicate-based access control (PBAC), and a fine-grained based access control to Swift object storage of the OpenStack cloud platform. PB-FGAC provides fine-grained access control to a predicate, which is part of the object, and helps to avoid access to the whole object. The user requested the OpenStack cloud receives by the NOVA component (NOVA is responsible to compute instances) to process the request, and the request is then forwarded to the object attribute storage services and a policy engine service, respectively. Each policy consists of object-level access and user-level access. After this, the PBAC service helps the user to access the specific data or predicate required. The results of the study demonstrate that the model provides more restricted access control than other environments with a default access control policy, such as the Amazon Web Services (AWS), Microsoft Azure, and Open Stack cloud platforms. However, the model proposed only applies to the JavaScript Object Notation (JSON) document, and its confidentiality also required addressing.

Finally, Ghaffar et al. [19] discuss the security issues associated with the data management operations in centralized cloud storage, such as insider attack, the lack of a data access verification model, and the lack of a sharing data configuration. They propose a modified model for data access and sharing in cloud storage, using a proxy key protocol. This model involves three entities, the DO, the DU, and the CS, together with three phases, namely data access, data storage, and a data sharing system. The data access involves the user and cloud server's authentication mechanism, and these then communicates with each other by sharing the session key. The data storage provides the users with the storage services required to share encrypted files or data with other desirable users. The data-sharing system searches by the user's keywords for the encrypted file after the user is authenticated. The method proposed is secure against multiple attacks, such as user or cloud impersonation and man-in-the-middle attacks, and data confidentiality breaches. However, the DU could search for a specific file by entering related keywords, resulting in the retrieval of multiple files, or a long search time.

Table I compares the centralized access control models, based on different criteria. The data confidentiality ensures the data is not a disclosure by unauthorized users. The scalability ensures when the number of users is increased, the system performs well.

B. Decentralized Access Control

In the decentralized access control model, there are four entities: the DO, the distributed center, the CS, and the DU, as shown in Fig. 3. A distributed center is a trusted authorization database that manages the access policy, and generates the keys for the cloud, the DU, and the DO. The distributed center can be a blockchain, a distributed CA, or a distributed attribute authority (AA). The DO uploads resources and stores the data to the CS, and publishes the access rights to the distributed server. Meanwhile, the CS stores the data, provides the authentication, and determines the permission of the DU and the DO. Finally, the DU receives and downloads the required data or resource, according to the access policy.

Al-Dahhan et al. [20] propose a distributed multi-authority CP-ABE. The system propose contains six entities and three phases. It commenced with an initialization phase, in which the CA must register each authorized DU and AA to obtain their identities. The DO creates an encrypted access control policy for the DU. Then, the DO publishes the encrypted access policy to the CS. One of the distributed AAs in the system then generates the keys for the DU to decrypt the access policy. The DU could then access and decrypt the resource, if they are authorized and matched the attribute assigned to him. If the user revokes his authorization, the CA informs the proxy server responsible for maintaining and updating the DU authorization list. The system proposed protects data confidentiality and secured against collusion attack. However, the DU performs a lot of computation for computing the secret key from one of the distributed AA.

TABLE I. A COMPARISON OF THE EXISTING MODELS FOR CENTRALIZED ACCESS CONTROL IN CLOUD COMPUTING

The proposed model	**Access Control**		**Data confidentiality**	**Scalability**
	Traditional Access Control	***Encryption- based***		
Online/offline CP-ABE scheme [14]	X	✓	✓	X
PriGuarder [15]	X	✓	✓	X
EFH-CP-ABE [16]	X	✓	✓	X
Agent-based ABE access control [17]	X	✓	✓	X
PB-FGAC [18]	✓	X	X	X
Ghaffar et al. [19]	X	✓	✓	X

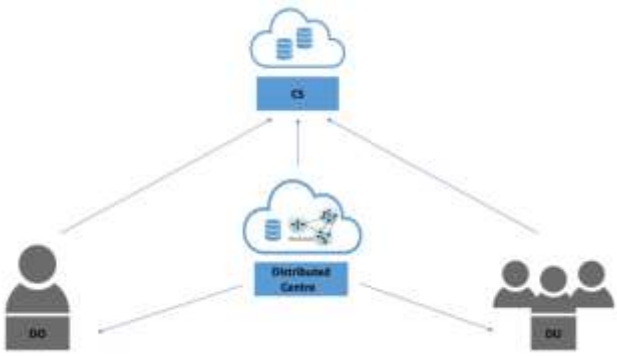


Fig. 3. Decentralized Access Control Model.

In another study, Wei et al. [21] propose a multiauthority CP-ABE to protect the outsourced data in the cloud. The approach proposed consists of five entities, each of which had a role. First, the third party produces a global parameter for the approach proposed. Second, each AA is responsible for generating a public parameter and the secret master key, and for managing the validity of each user's attribute belong to its area. Third, the DO defines its access policy on the DU attribute, and sends the encrypted data to the cloud. Fourth, every DU has a unique global identifier, set of attributes, and secret key. Finally, the CS is responsible for storing the data and updating any information requested from any entity. The proposed approach is secure against collusion attacks. However, the number of attributes has an influence on the efficiency of the scheme's algorithm.

Wang et al. [22] propose an integrated technique that combined blockchain, smart contracts, and CP-ABE. The model proposed consists of four entities, the DO, the DU, the CS, and blockchain, and involved four phases, namely system initialization, file encryption, key generation, and file decryption. The DO is responsible for defining the access control policies, determining the attribute sets, and creating the smart contract with valid access periods to the DU. Meanwhile, the DU is required to satisfy the encrypted smart contract's attribute set to decrypt it, then to obtain the content key to decrypt the stored encrypted files, and access them in the cloud. The CS is responsible for storing the encrypted resources or files, and finally blockchain is responsible for deploying and storing the smart contracts. The cost of the data access is low, and the performance is feasible. However, the study does not consider the integrity of the file uploaded by the DO.

In another study, Yang et al. [8] propose a mechanism named Authprivacychain, which uses the blockchain technique, merged with smart contracts. The model proposed consists of four entities, the DO, the DU, the CS, and blockchain., It involves four phases, namely initialization, access control, authorization, and authorization revocation. First, the DO uploads the resources required to the cloud and registered the transaction in the blockchain, then publishes the authorization to the DU. The DU then sends a request to the CS for a specific resource, and the CS sends a query to the blockchain, and evaluates whether the resource requested by the DU has permission or not. Finally, the CS replies to the request with the stored access. These processes are designed in an encrypted manner to secure data privacy. The approach proposed is secure against external and internal attacks, and the Authprivacychain protects the confidentiality, integrity, and accountability of the resources, as well as the availability and authenticity. However, the time performance of the technique proposed depends on the blockchain node configuration. This is because there are different types of the blockchain with various configuration parameters. For example, block time generation, a block take two minutes while other take 10 mintues [9].

The difference between [8] and [22] is in the decryption of the resource. The DU could access the resources once they are authorized by the DO, and the decryption key is delivered at the initialization phase in [8], but in [22] the DU is required to satisfy the attribute set by the DO, in order to obtain the decryption key of the encrypted resource.

Meanwhile, the authors of [23] propose a decentralized attribute ABE access control model for a mobile cloud. The scheme proposed is similar to that in [18], but the algorithm specification differed. The proposed approach is secured against reply and collusion attacks. However, it fails to achieve a public ciphertext test. Furthermore, the computational complexity increases with the attribute number involved in the ciphertext.

Table II compares the decentralized access control models, including the different key features used in the access control models, namely the access control permission, the authorization, the authorization revocation, the authentication identity, and the access log. In terms of the access control permission, it specifies what right the user has to the data, such as read or write. Meanwhile, in terms of authorization, it specifies the permission required for a user to access the cloud resource or data. The authorization revocation refers to the authorized users' ability to revoke, dispense, or delete the resource access permission of other permitted users. The authentication is also the mechanism used to verify that the user is who they claim to be. Finally, the access log contains all the information of all the requests for user actions on the cloud resource. Table III summarizes the advantages, limitations, and type of architecture of the existing access control models in cloud computing.

TABLE II. A COMPARISON OF THE EXISTING MODELS FOR DECENTRALIZED ACCESS CONTROL IN CLOUD COMPUTING

The proposed model	Techniques	Access control permission	Authorization	Authorization revocation	Authentication identity	Access log
Al-Dahhan et al. [20]	Multi-authority CP-ABE	✓	✓	✓	A symmetric key	X
Wei et al. [21]	Multi- authorityABE	✓	✓	✓	Global identifier	X
Wang et al. [22]	blockchain , smart contract , CP-ABE	✓	X	X	blockchain wallet address	✓
Authprivacychain [8]	blockchain , smart contract	✓	✓	✓	blockchain wallet address	✓
De et al. [23]	Multi- authorityABE	✓	✓	✓	Global identifier	X

TABLE III. A COMPARISON OF ACCESS CONTROL MODELS IN CLOUD COMPUTING

Year	Access Control Model	Architecture	Advantages	Limitations
2018	Online/offline CP-ABE [14]	Centralized	Improved the computational overheads of the current schemes for EHR.	Trade-off between the computation time and storage space required addressing.
	PriGuarder [15]	Centralized	The key power in PriGuarder was the AFG method, which protects user privacy.	A malicious TTP may disclosure users' privacy.
	Al-Dahhan et al. [20]	Decentralized	Protected data confidentiality and secured against collusion attacks.	The computation cost for computing the secret key generation from the AA was the responsibility of the DU.
	Wei et al. [21]	Decentralized	Secure against collusion attacks.	The number of attributes had an influence on the efficiency of the scheme's algorithm.
2019	EFH-CP-ABE [16]	Centralized	Encrypted multiple files at a similar access level.	Computation time for the encryption and decryption required enhancement.
	Agent-based ABE access control [17]	Centralized	Secure against various attacks, such as malware injection, identity theft, collusion, and certification authority failure attacks. The response time for reading the data access in the cloud was improved.	The process of selecting the appropriate backup authority node required securing.
	Wang et al. [22]	Decentralized	Low cost of the data access, and feasible performance.	Did not consider the integrity of the file uploaded by the data owner.
2020	PB-FGAC [18]	Centralized	Provided partial access to the JSON document.	Only applied to the JSON document. Confidentiality also required addressing.
	Ghaffar et al. [19]	Centralized	Secure against multiple attacks, such as user or cloud impersonation, and man-in-the-middle attacks. Provided data confidentiality.	Searching by DU for a specific file resulted in the retrieval of multiple files, or wasting time.
	Authprivacychain [8]	Decentralized	Secure against internal and external attacks. Protected the confidentiality, integrity, and accountability of the resources, as well as availability and authenticity.	The performance depended on the blockchain node configuration.
	De et al. [23]	Decentralized	Secure against reply and collusion attacks.	Failed to achieve the public ciphertext test. The computational complexity increased with the attribute number involved in the ciphertext.

IV. CHALLENGES AND FUTURE RESEARCH DIRECTION FOR ACCESS CONTROL MODELS IN CLOUD COMPUTING

Section III discussed the different solutions currently proposed for access control models in cloud computing. However, certain issues and limitations remain that must be addressed by future studies. This section presents the challenges and suggested future research direction for both centralized and decentralized access control models in cloud computing.

A. Centralized Access Control

Section III-A presented the current solutions based on centralized access control, but the studies discussed relied on either traditional access policy models or encryption based models, both of which are insufficient for securing a system. More hybrid works that combine both access policy models and encryption models are required to ensure confidentiality and integrity.

Importantly, the studies in centralized access control assume that a CA or an AA is trusted. While, the CA or AA is not always trusted, which causes a number of issues, such as leakage of sensitive data and keys. Thus, it leads to disclosure users' privacy. Also, the CA may suffer from failure for any reasons such as attacks but the current process [17] of selecting the appropriate backup authority node required securing.

Moreover, in terms of centralized cloud systems, the extant studies employed multiple techniques and algorithms, such as public-key encryption, symmetric encryption, and hash algorithms. These studies store the resources in an encrypted manner in the cloud storage, however, searching a particular resource or a specific file by entering related keywords [19], resulting in the retrieval of multiple files, may produce erroneous outcome, or an extended search time.

Furthermore, the computation time for the encryption and decryption is a concern which needs to be enhanced when encrypting multiple files at a similar access level. However, in mobile cloud computing, trade-off between the computation time for the encryption and decryption and storage space required addressing.

Finally, the number of users in the cloud system is increased day by day; thus more attention is needed for scalability to ensure the system's reliability.

B. Decentralized Access Control

Currently, the concept of adopting distributed authority to tackle the issue of centralized authority is ongoing, such as the work of [8, 22, 23]. However, limitations remain that must be addressed, such as the restoration of data from a compromised AA or CA.

The blockchain (decentralized) uses a hash function to generate a unique id for each resource and to store it in a smart contract, which means that the unique id provides a quick search with privacy, and returns the correct result [13]. Therefore, the reading requests in decentralized cloud models are faster than in centralized cloud models. There is a need for more focus on the application and development of blockchain technology and smart contracts to overcome their existing limitations, such as the blockchain node configuration, slow performance in writing access rights, blockchain node registration, and computational resources.

Furthermore, the authorization revocation discussed in Section III-B requires different reasons to perform, such as a user sending a request for the revocation. Also, the resource may be compromised, and the cloud server may therefore wish to disable it. The process of revocation therefore requires more attention to ensure that it is secure and fast [8].

Finally, the access log contains records of all the entities' interactions in the system. It is of essential value, as it can provide a mechanism to locate the interaction responsible for an attack on the system. Moreover, security analysis can perform post-audit analysis on these records to detect any vulnerabilities in the system. Consequently, there is a need to involve access log design within the system implementation [8].

V. Conclusion

Cloud computing provides different services via the Internet that may engender a number of security issues, such as unauthorized access to cloud resources. Access control is a security technique that controls the access to cloud services. This paper provided a taxonomy for access control models in cloud computing, according to centralized and decentralized models, and discussed a range of works that employed this taxonomy, comparing the advantages and limitations of each.

The study concluded that, in order to improve the security of access control models in cloud computing, there is a need for different solutions, both centralized and decentralized. In centralized access models, the system should ensure both confidentiality and integrity, and tackle the issue of long search time. While in decentralized access models, authorization revocation must be faster and more secure. Moreover, the use of access logs to keep track of any disclosure of the system is required, and blockchain node configuration must be enhanced to improve the system's performance. The study also presented the current challenges and recommended a direction for future research in access control models in cloud computing.

References

[1] K. Albulayhi, A. Abuhussein, F. Alsubaei, and F. T. Sheldon, "Fine-Grained Access Control in the Era of Cloud Computing: An Analytical Review," in 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), 2020, pp. 0748-0755.

[2] V. Winkler, "Chapter 2 - Cloud Computing Architecture," in Securing the Cloud, V. Winkler, Ed. Boston: Syngress, 2011, pp. 29-53.

[3] V. Winkler, "Chapter 1 - Introduction to Cloud Computing and Security," in Securing the Cloud, V. Winkler, Ed. Boston: Syngress, 2011, pp. 1-27.

[4] P. J. Kaur and S. Kaushal, "Security concerns in cloud computing," in International Conference on High Performance Architecture and Grid Computing, 2011, pp. 103-112: Springer.

[5] S. Curry et al., "Infrastructure security: Getting to the bottom of compliance in the cloud.," The Security Division of EMC (2010).

[6] F. Cai, N. Zhu, J. He, P. Mu, W. Li, and Y. Yu, "Survey of access control models and technologies for cloud computing," Cluster Computing, vol. 22, no. 3, pp. 6111-6122, 2019/05/01 2019.

[7] S. Xie, Z. Zheng, W. Chen, J. Wu, H.-N. Dai, and M. Imran, "Blockchain for cloud exchange: A survey," Computers & Electrical Engineering, vol. 81, p. 106526, 2020/01/01/ 2020.

[8] C. Yang, L. Tan, N. Shi, B. Xu, Y. Cao, and K. Yu, "AuthPrivacyChain: A Blockchain-Based Access Control Framework With Privacy Protection in Cloud," IEEE Access, vol. 8, pp. 70604-70615, 2020.

[9] M. Schäffer, M. Di Angelo, and G. Salzer, "Performance and Scalability of Private Ethereum Blockchains," 2019, pp. 103-118.

[10] W. Stallings, L. Brown, M. D. Bauer, and A. K. Bhattacharjee, "Chapter 4 - Access Control " in Computer security: principles and practice: Pearson Education Upper Saddle River, NJ, USA, 2012, pp. 113-154.

[11] V. Winkler, "Chapter 5 - Secure the Cloud: Data Security," in Securing the Cloud, V. Winkler, Ed. Boston: Syngress, 2011, pp. 125-151.

[12] S. Pavithra, S. Ramya, and S. Prathibha, "A Survey On Cloud Security Issues And Blockchain," in 2019 3rd International Conference on Computing and Communications Technologies (ICCCT), 2019, pp. 136-140.

[13] S. Wang, Y. Zhang, and Y. Zhang, "A Blockchain-Based Framework for Data Sharing With Fine-Grained Access Control in Decentralized Storage Systems," IEEE Access, vol. 6, pp. 38437-38450, 2018.

[14] Y. Liu, Y. Zhang, J. Ling, and Z. Liu, "Secure and fine-grained access control on e-healthcare records in mobile cloud computing," Future Generation Computer Systems, vol. 78, pp. 1020-1026, 2018/01/01/ 2018.

[15] L. Lin, T. Liu, S. Li, C. M. S. Magurawalage, and S. Tu, "PriGuarder: A Privacy-Aware Access Control Approach Based on Attribute Fuzzy Grouping in Cloud Environments," IEEE Access, vol. 6, pp. 1882-1893, 2018.

[16] J. Li, N. Chen, and Y. Zhang, "Extended File Hierarchy Access Control Scheme with Attribute Based Encryption in Cloud Computing," IEEE Transactions on Emerging Topics in Computing, pp. 1-1, 2019.

[17] F. Jamal, M. T. Abdullah, Z. M. Hanapi, and A. Abdullah, "Reliable Access Control for Mobile Cloud Computing (MCC) With Cache-Aware Scheduling," IEEE Access, vol. 7, pp. 165155-165165, 2019.

[18] C. Anilkumar and S. Subramanian, "A novel predicate based access control scheme for cloud environment using open stack swift storage," Peer-to-Peer Networking and Applications, 2020/07/26 2020.

[19] Z. Ghaffar, S. Ahmed, K. Mahmood, S. H. Islam, M. M. Hassan, and G. Fortino, "An Improved Authentication Scheme for Remote Data Access and Sharing Over Cloud Storage in Cyber-Physical-Social-Systems," IEEE Access, vol. 8, pp. 47144-47160, 2020.

[20] R. R. Al-Dahhan, Q. Shi, G. M. Lee, and K. Kifayat, "Revocable, Decentralized Multi-Authority Access Control System," in 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion), 2018, pp. 220-225.

[21] J. Wei, W. Liu, and X. Hu, "Secure and Efficient Attribute-Based Access Control for Multiauthority Cloud Storage," IEEE Systems Journal, vol. 12, no. 2, pp. 1731-1742, 2018.

[22] S. Wang, X. Wang, and Y. Zhang, "A Secure Cloud Storage Framework With Access Control Based on Blockchain," IEEE Access, vol. 7, pp. 112713-112725, 2019.

[23] S. J. De and S. Ruj, "Efficient Decentralized Attribute Based Access Control for Mobile Clouds," IEEE Transactions on Cloud Computing, vol. 8, no. 1, pp. 124-137, 2020.

An Efficient Color LED Driver based on Self-Configuration Current Mirror Circuit

Shaheer Shaida Durrani[1], Abu Zaharin Bin Ahmad[2], Bakri Bin Hassan[3], Atif Sardar Khan[4], Asif Nawaz[5], Naveed Jan[6]
Rehan Ali Khan[7], Rohi Tariq[8], Ahmed Ali Shah[9], Tariq Bashir[10], Zia Ullah Khan[11], Sheeraz Ahmed[12]
Faculty of Electrical & Electronics Engineering Technology, University Malaysia Pahang, Pekan, Malaysia[1, 2, 3]
US-Pak Centre for Advanced Studies, In Energy (USPCAS-E), UET Peshawar, Pakistan[4]
Faculty of Electronics, Higher Colleges of Technology, Dubai, UAE[5]
Department of Information Engg Tech, University of Technology, Nowshera, Pakistan[6]
Department of Electrical Engineering, University of Science and Technology, Bannu, Pakistan[7]
School of Information and Technology (SIT), King Mongkut's University of Technology, Thonburi, 10140, Bangkok, Thailand[8]
Department of Electrical Engineering, Sukkur IBA University, Sukkur, Pakistan[9]
Department of Electrical Engineering, COMSATS University, Islamabad, Pakistan[10]
Manager Network Systems, Directorate of Science and Technology, Government of KPK, Pakistan[11]
Department of Computer Science, Iqra National University, Peshawar, Pakistan[12]

***Abstract*—The string channel of Color LED driver with precise current balancing is proposed. It is noted that to drive a multiple LEDs string is by using a proper current source, due to the level of the brightness LED depends on the quantity of the current flows. In the production of LEDs, the variation in the forward voltage for each LED has been found significantly high. This variation causes different levels of brightness in LEDs. Then, controlling load current of LED by using a resistor to limit the LED current flowing is considered by associated with the forward voltage, instantly. Current sources have been designed to become immune to the above problem since it regulates the current, and not the voltage which flows through the LEDs. Hence, constant current source is the essential requirement to drive the LEDs. Besides, it is complex for color LEDs, dependent on the number of control nodes and dimming configuration. To arrange an accurate load current for the different sets of string color LEDs, the efficient LED driver is required, in which the current sharing is complement to each LED strings. Therefore, this paper suggests a color LED driver with self-configuration of enhanced current mirrors in multiple LED strings. The load currents have been efficiently balanced among the identical loads and unequal loads. The comparable efficiency of the string color LEDs losses has been shown thoroughly.**

Keywords*—Color LED driver; current mirror circuit; super diode

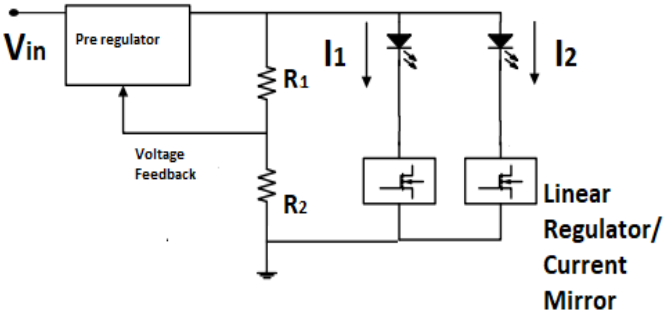
I. INTRODUCTION

Nowadays, light-emitting diode (LEDs) has gradually replaced the lighting system due to better luminous, affordable size, energy-saving, and nature friendly [1]. Nonetheless, the constant output current for reliable LEDs driver is vital to ensure a good performance of LEDs. For color LEDs, the balancing for current sharing through the string of red, green, or blue is crucial for avoiding blackout from a single LED fault due to an excessive driving voltage; hence the parallel string arrangement is preferred. Besides, it works with the LED driver to provide equivalent current to each string to make it sure to have uniform brightness. The inherent imbalance current flow could occur due to the LED forward voltage spread. Parallel channel connections of the LEDs create current differences among the LED strings due to their characteristic deviations. These deviations reduce the life cycles of the LEDs by introducing thermal spotting at a particular point. They also raise the matter of non- uniform luminance from the LEDs. To increase efficiency and simplicity, multichannel LED drivers are developed to replace single-channel LED drivers. Additionally, multichannel LED drivers can operate at different brightness for each individual LED channel. This enables lighting applications to be more optimized as compared to the use of single-channel LED drivers. Hence, this leads to advancements of multichannel selective dimming LED drivers. Parallel strings of the LEDs are frequently used in various embellishing, lighting, illustrating, and signaling purposes. The reliability of these circuits contains a significant factor, specifically in the field of backlighting system. Similar current values in each LED string are a crucial considered factor since it affects the reliability of the whole circuit. A small difference in current values in the strings has the potential to adversely affect the lifetime reliability and time span of the circuit. Since it is impractical to create identical devices, current balancing techniques have enormous importance for proper functioning of LED arrays. It has been observed that the most effective way of dealing with current imbalance is the utilization of current mirror (CM) techniques.

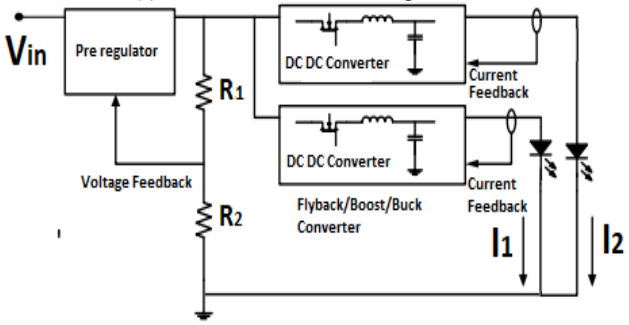
Up to date, most of the drivers use a straightforward method as a common control supply to run LEDs which driven with independent sources of constant current [2]. In practical, color LEDs has a complex nature in a LED backlighting system, which depends upon the number of control nodes and LEDs requirement [2]. A switch-mode based power converter (SMPC) has been used for providing a steady current source to drive color LEDs. Nevertheless, it suffered from an intemperate control scattering in its series-pass devices [3]. A different utilization of driving color LEDs with numerous strings is being actualized for background brightening applications of LCD [3]. In this approach,

numerous current controllers have been utilized. As a result, significant improvement has been seen recently, for white power and high brightness (HB) LED technology [1], which made the usage of LED lighting sources to become prominent in most display application. Since the LED brightness coordinating current than voltage, the current balancing strategies which equalize the current through the string are vital [4]. Besides, the V-I characteristic of LEDs showing a considerable current change if a small diversity occurs in the voltage source to the LED. Since the color and brightness of LEDs are relying on the load current, hence, to avoid color shifting and to maintain the brightness concentration, the operations of color LEDs are recommended to be organized with a constant steady current. As in [5] and [6], demand always exists for compact and straightforward LED and color LED drivers. Nonetheless, current balancing of the color string LEDs contributes to the overall efficacy of the module LED loads or LED luminaires, rendering to maintain color produce and LEDs aging. Therefore, [7] had proposed the current mirror (CM) in solving the current imbalances among the LED strings with omitting an auxiliary source supply. It has shown significant improvement in maintaining the current balance in the string. However, separate source of supply to regulate the LEDs is required, which incline to the limitation and drawback of the existing CM. In addition, the LEDs display need to be turned on and off quicker than other lighting devices, hence the most appropriate current sharing circuit for the fast switching is desired. Therefore, the improvement circuit arrangement of CM has been suggested in this paper, which is also concerning the power dissipation through the string. Due to certain problems in the manufacturing process, LEDs have a problem with relatively large variations in the forward voltage (FV) characteristics [8]. These variables generate current-sharing problems, which differences between the load currents in rows of LEDs connected in a parallel fashion. This result does not permit the uniform distribution of the heat in the lighting, thus accelerating the aging of specific LEDs and getting luminance in a non-uniform manner. Eventually, the quality and the reliability of illumination of LED lighting devices would degrade. To resolve this problem, some suggestions have been suggested. The solutions can be classified into the method of using a linear regulator and CM in each row, converter for current control in each row, and by compensating the current error using a passive device in each row as discussed in Fig. 1(a) and (b). The balancing transformer is used in such applications for discharge lamps (CCFL, fluorescent lamps, etc.) as shown in Fig. 1(c). These lamps are run with the help of AC source, so a balancing transformer is directly applicable. However, for LEDs run through a DC source, the redesign of the circuit is necessary.

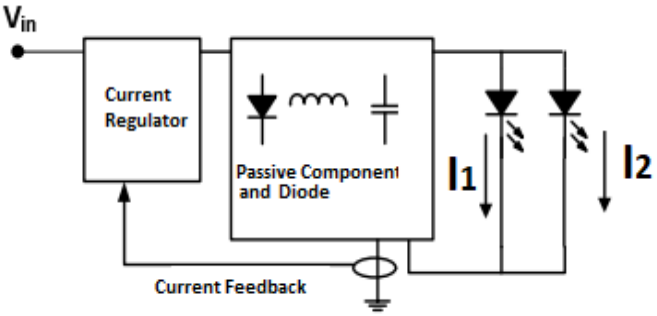
Therefore, in this article, a driving circuitry for parallel-connected color LEDs is presented, with the aiding of the current mirroring circuit. In this method, the suggested circuit can be reduced in size as compared to the conventional available design circuit. As a scope, a color LED driver has been proposed to drive 9 parallel-connected LEDs, consisting of either red, blue or green LEDs. The put forward driver has been designed to maintain the rock bottom drive voltage across the LEDs and its associated transistor (current controller); leading to reduced power dissipation across the transistor, which eventually increased the efficiency of the particular LED string or set of LED strings. The suggested system of color LED driver is dimming capable for each individual LED through the controller circuit. Efficiencies have been verified at 97%, 98.55% for the red and green/blue, at the 21 mA and 22 mA, respectively. This paper starts with the discussion of limitations in the existing CM and then discussing the proposed CM circuit in section three. The rest follows the discussion of results and analysis.



(a) Current Mirror of Linear Regulator [3].



(b) DC/DC Converters [4].



(c) Passive Components and Diodes [8].

Fig. 1. Methods of Linear Regulator.

II. MOTIVATION

The existing CM approach needs a separate source of supply for regulating current at the LEDs as loads or in other words needs one fixed current source as a reference from separate power. In this approach, the reference is limited to the single load current only and the inability to allow predetermine the variations of the LED strings. As in [9 -13], traditional CM has a problem regulating loads strings of LEDs, while operating controlled current from low supply voltages. It becomes difficult to operate LEDs with minimal voltage, which results in operational demand for overhead

voltage. The operational demand is required in handling the variations of the forward voltage drop across the loads, especially in mass-produced LED modules. Without the headroom voltage source, it becomes difficult to get equal currents parallel to the strings of LED loads. To address such limitations, a CM that works in a self-configurable mode has been proposed in [10]. The circuit's operation depends upon a Darlington pair transistor, which has the ability to set the minimum current as the current reference to each string by closed-loop current control configuration. The Darlington transistors work in a linear region, to counter the differences of voltages between the bus of the dc voltage and the string voltage and do not need headroom voltage apparently, which suitable for LED backlight [11]. However, the driving currents, i.e. the base currents generating from the reference to bias the transistors, make the current replication imperfect even in the well-matched case which causes accuracy errors. This error increases with the number of strings, whereas its operation can be described with the following equation.

$$I_{O1} = \frac{\beta}{\beta+N} I_{REF} \quad (1)$$

Where N shows the number of LED strings and I_{REF} denotes for target current. Since the LED current is almost corresponding to the LED's illumination, it is crucial to control the current precisely. In the event that all the LEDs are running in a single string, in series, there is no issue of mismatching since each LED has a similar current level. As the number of LEDs being used increases, resembling strings gets important, and a decision must be made with regards to how to control the current in each string. LEDs makers use binning to sort parts into bunches that precisely coordinate the LEDs forward voltage drops to permit execution. Conventionally, a fixed-voltage source and adding a straightforward resistor to set the current level has been done, but it costs to the efficiency drop. The loss of energy causes the output response of the LED to become slower. Secondly, the forward drop of voltages of the LEDs decreases with the increase of temperature. If somehow one channel or string gets significantly hotter due to any internal or external disturbing factor, its forward drop suddenly reduces and starts to draw huge current. This led to power dissipation in the form of heat. If disturbing factor is not removed, current will keep increasing and possibly lead LED to fail. Such situation needs that the applied voltage for driving currents in the strings is kept regulated. Thirdly, if an LED becomes open-circuited in the regulated string, the applied voltage energizing the strings is controlled by the control circuit and eventually causes overvoltage in the unregulated string, leading to failure. So, appropriate design is needed to avoid such issues. Furthermore, dimming is important factor in the effective control of lighting and in saving energy, but it faces various challenges. Getting full range of dimming means complete control of current passing through the LED. There are ways to energize color LEDs in the module system and the easiest way is to use respective constant current sources for each load i.e. string of LED or an LED [1]. This method looks easy, but it is costly and needs more components to the driver circuitry, which creates the whole system complex. Adding components simply means adding the possibilities of failure modes.

III. Proposed Current Mirror (CM)

Two LED strings (two parallel-connected of color LEDs) are presented for ease of understanding for the proposed CM. The circuit can be extended to the number of strings. The buffer amplifier circuit with small magnitude is imposed as feedback requirement to replace the diode. The voltage drops across a small resistor decrease to the forward voltage drop, which is getting nearly zero when divided with the op-amp's open-loop gain. The feedback mechanism of the op-amp has a feature of making voltage collector of transistor of the CM circuit equal to the voltage at the base, hence preventing the saturation from transistor. A negligible offset voltage occurs across the resistor. This condition is useful in improving CM's circuit operation. Whereas the insignificant increase in the LED forward voltage at each string, creates a reference current for the entire proposed CM circuit. This proposed CM circuit is modular in nature and comprises of two main circuits, which are emitter-coupled logic (ECL) and super diode circuits are depicted in Fig. 2. In this approach, the circuit allows the CM to take the string's current as a reference current, in which the maximum voltage drop occurs in the string, as translated in Fig. 3. Furthermore, the resistors are series wise added at the outputs of the super diode circuits, which are further connected in series wise with the ECL transistors' bases. In the conventional CM circuit, the transistor in each string operates like a couple of conjugated diodes which the forward biasing is required to turn on the diode due to base-emitter connection to activate the LEDs load on. It requires a minimum 0.6 V to activate which eventually increasing power loss. In that case, the enhancement of the CM by modification of the circuit is necessary. In the proposed CM circuit, the power losses are comparatively less than the improved Wilson CM circuit because the BJTs are not connected series wise in each string [5]. In implementation, the proposed CM circuit requires a closed control for superior execution, in which the DC-DC converter is utilized for feedback signal.

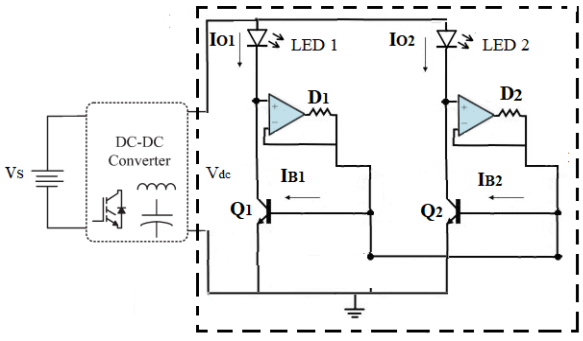


Fig. 2. Combinational Circuit of Super-Diode and ECL Circuit.

IV. Mathematical Analysis

In evaluating the performances, the mathematical analysis is carried out. For the sake of simplicity analysis, only two leg strings are discussed. Refer to Fig. 2, suppose that LED string 1 has the lowest current branch, such that V_{CE1} is lower than V_{CE2}. This phenomenon turns the super diode D_1 on, and the associated op-amp feedback makes the Q_1 transistor's

collector to has the same potential as the potential available for the transistor at its base, and the source of the supply voltage is maintained as;

$$V_S = V_{CE} + V_{FORWARD} \quad (2)$$

The supply voltage does not have any headroom voltage element due to the nature of the ECL circuit, then the betas of the transistors are considered equal to each other such that;

$$\beta_1 = \beta_2 = \beta \; ; \; I_{\beta 1} = I_{\beta 2} = I_\beta \quad (3)$$

Since the current goes through the LED string, connected to the proposed CM circuits are equal to each other, hence.

$$I_{o1} = I_{o2} = \beta I_\beta \quad (4)$$

Finally, the power dissipation in each LED is represented as

$$P_{LED} = R_{LED\_LOAD} \times I_{On} \quad (5)$$

Where, R_{LED_LOAD} is a resistive load of the LEDs, I_{on} is the load current in the *n*th string, V_s denotes the supply voltage and $V_{FORWARD}$ is a LED forward voltage. As the P_{S_STRING} is equal to the total of all the power developed across all the resistances of the devices in series, thus, the efficiency (η) could be computed as follows.

$$\eta = \frac{P_{LED}}{P_{S\_STRING}} \times 100 \quad (6)$$

where, P_{S_STRING} is accumulative P_{LED} and losses across the transistor. It has proved that the proposed circuit places no additional headroom voltage other than VCE or VBE during operating, the base of transistor becomes short and hence virtually a part of the transistor's collector. Thus, the only power loss occurs in the string, develops across the collector and the emitter terminal of the transistor. This helps in having a proper biasing of the transistor circuit. As a result, the power loss across the transistor could be reduced to a significant extend.

The following readings have been noted experimentally, at 22.0 mA current for the red LED. In which, P_{LED} and P_{S_STRING} are computed 45.0 mW and 1.10 mW, respectively. Thus, efficiency can be calculated as in the following.

$$\eta = \frac{P_{LED}}{P_{S\_STRING}} \times 100 = 97\% \quad (7)$$

Similar reading and calculations have been obtained for the blue LEDs in addition to the green LED. Since the blue LEDs besides the green LEDs have the same resistive characteristic, the reading is computed together at 21.0 mA. The P_{LED} and P_{S_STRING} are calculated up to 69.62 mW and 1.02 mW, respectively. Hence, the efficiency is calculated to 98.55%.

V. Dimming Control Mechanism

By obtaining the proper biasing, preventing the operation transistor to saturation, or in cut-off mode. This matter is engaged to the proper arrangement of dc collector current at a certain dc voltage by setting up a proper quiescent point as the circuit as discussed with the help of Fig. 3. At first, a base resistor (R_B) is rearranged in the middle of collector and base of the transistor due to the nature of the circuit does not allow base resistor connected to the supply. Hence, the connection between base and LEDs load is opted by tapped the shunt node in the middle of LEDs and collector transistors. Then, resistor R_1 is located in the middle of the base and emitter for better biasing events which provided from the voltage developed across the R_1.

Virtual resistance is a kind of resistance that does not present physically in the middle of the dimming circuit and the biasing circuit of the transistor. The dimming circuit creates various ground voltage references for the flow of load current from the biased transistor. Thus, by varying different voltage ground references, it is possible to create a kind of so-called virtual resistance in the path of flow of load current, in which no physical resistance is required. However, in this approach, the color LEDs module consists of a set of 9 color LEDs are configured, as shown in Fig. 4. Most of the LEDs dimming is controlled through PWM output by bringing change in the duty cycle. However, in this approach, the color LEDs module consist of a set of 9 color LEDs are configured. The supply comes from a DC-DC converter. Meanwhile, the dimming frequency is set to 1 kHz, where the dimming phenomenon is implemented at each leg of the CM circuits, in which the color LEDs are devised to be operated in individual dimming and overall diming. For ease computation, the power losses occur across the transistor are neglected.

As a result, the proposed CM associated with the dimming circuit possible to introduce a significant dimming level for single and whole LEDs in each string of color LEDs. While dimming occurs at certain LED loads, the associated transistor of the string reduces the current through the LED loads by raising the emitter's voltage of the transistor that working with the ECL circuit. It eventually rises the collector's voltage of the transistor which gives freedom to the LED load from the rest of the system which still running with a constant source of voltage. It is showing the proposed circuit topology could provide a constant load current with a dimming mechanism.

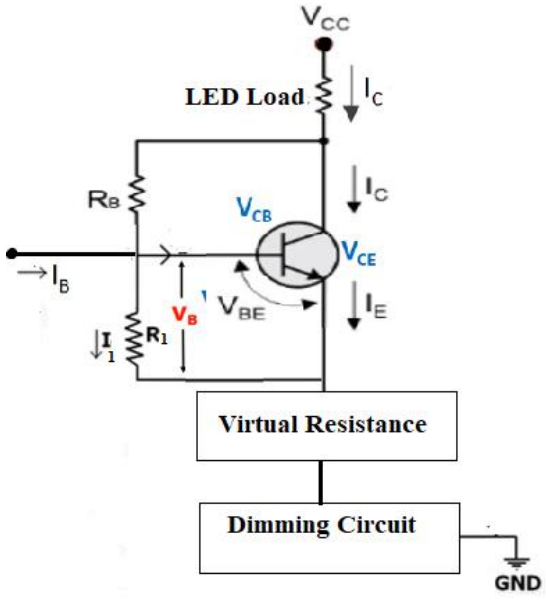


Fig. 3. Dimming Control Mechanism.

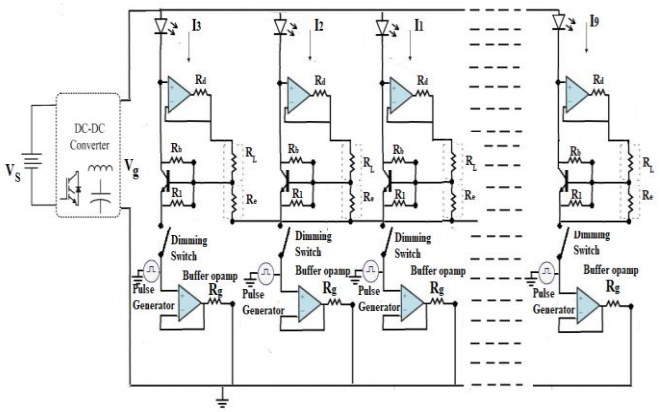


Fig. 4. Proposed CM with Dimming Circuits.

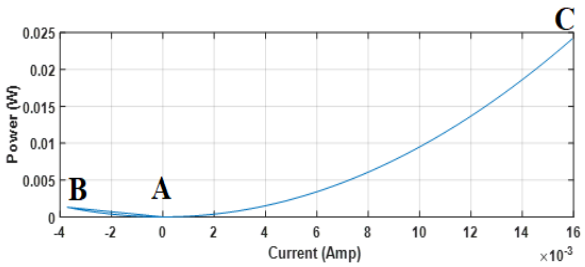


Fig. 5. Logarithmic Curve.

The switching device has been built to give the least resistance to load current. Its resistive value is very small. Hence the voltage drop across the load could be neglected. It gives rise to a voltage barrier to the flow of current. Because of it, the load current through the load has quadratic relationship with the power consumed by the load itself as shown in Fig. 5.

VI. Monte-Carlo Sensitivity Analysis

To evaluate the current balancing feature of the proposed driver system, a Monte Carlo procedure has been carried out, while introducing 2 small incremental resistive loads out of 9 loads of purely resistive nature. The statistical approach is carried out using a Monte-Carlo analysis to see the parameter performances. The resistive feature in the color LEDs is assumed to be random variables that have a two-dimensional normal distribution with truncations between the limits of lower and upper values of 90 Ω (R_1) and100 Ω (R_3), respectively, while keeping other 6 resistances of the system to 95 Ω (R_2). The simulation has been done on the proposed circuit while keeping the standard deviation of 1. Since the results of the remaining 6 strings of LEDs are identical to 95 Ω (R_2), hence only the results of three strings from proposed circuit are discussed which represented three strings of color LEDs.

The Monte-Carlo test has been schemed in Fig. 6 as well as in Fig. 7. The analysis has been done without and with a proposed current mirroring circuit. Furthermore, it has been done with the calculations for correlation of R_2 with other two resistors, i.e., R_1 and R_3. Further analysis also has been done with respect to their qualitative analysis of the probability occurrence changes of various resistive values. In Fig. 6(a), there is no current mirror circuit and the correlation between R_1 and R_2 dragging the R_1 towards the resistive value of 85 Ω. Where in Fig. 6(b), after applying the current mirror circuit, the resistive nature of R_1 is dragged towards 95 Ω, to catch up with the resistors of the system having 95 Ω, other than R_3. While inspecting R_1 with and without a current mirror (Fig. 6(c) and Fig. 6(d)), it has been noticed that at the base, the right-hand side of the bell-shaped (of current mirror circuit) curve looks wider (from the dotted center to the circles, circle pf the latter configuration shows very minute occurrence of probabilities) and has high probability of dragging the resistance, i.e., R_1 to 95 Ω resistive value. Then, in Fig. 7 same sort of discussion has also been observed in case of R_3 with and without current mirror circuits. In Fig. 8, it has also been found that there exists a probability of R_2 to change its resistive value to 100 Ω without the current mirror circuit and with the current mirror, the probability of having 100 Ω has been reduced significantly. It can be seen that their probability of occurrences has been squeezed to the region of 95 Ω.

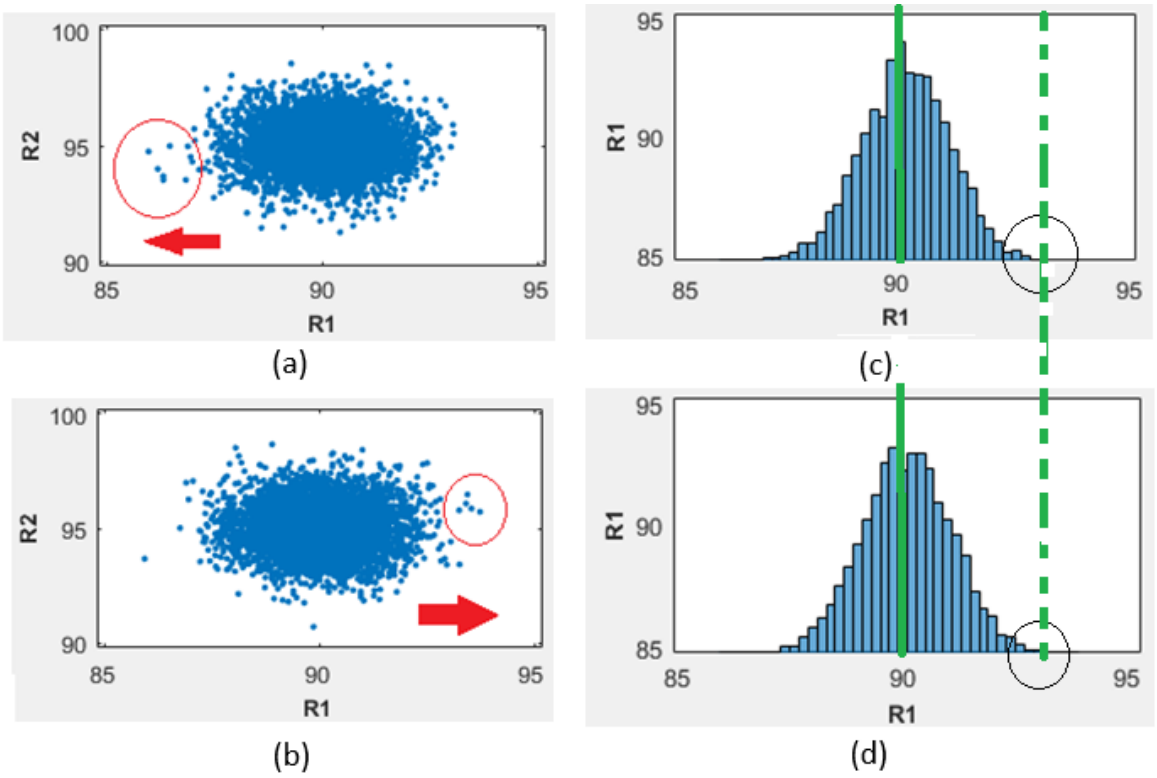


Fig. 6. (a)(b) Correlation between R_1 and R_2, without and with Current Mirror Circuits. (c)(d) Probabilities of Occurrences Change of Various Resistive Values

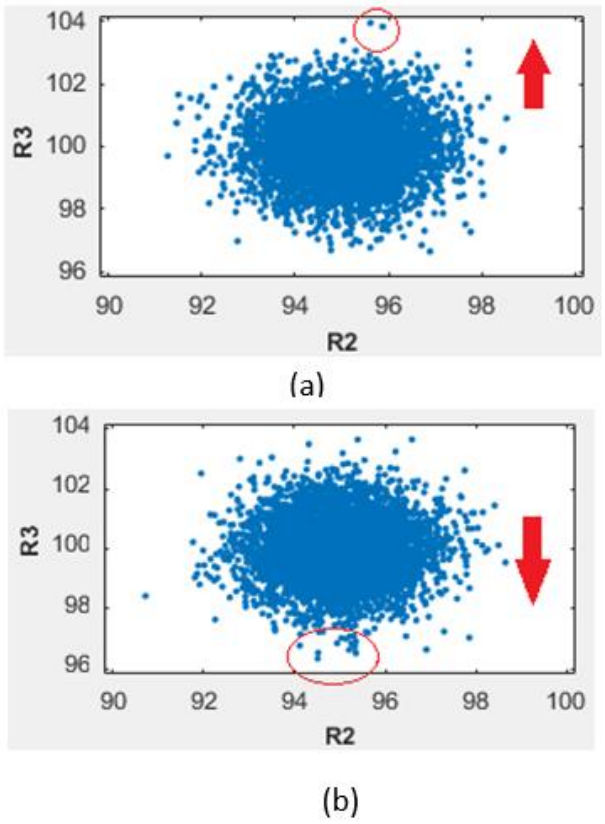


Fig. 7. (a)(b) Correlation between R_2 and R_3, without and with Current Mirror Circuits, (c)(d) Probabilities of Occurrences Change of Various Resistive Values of R_3.

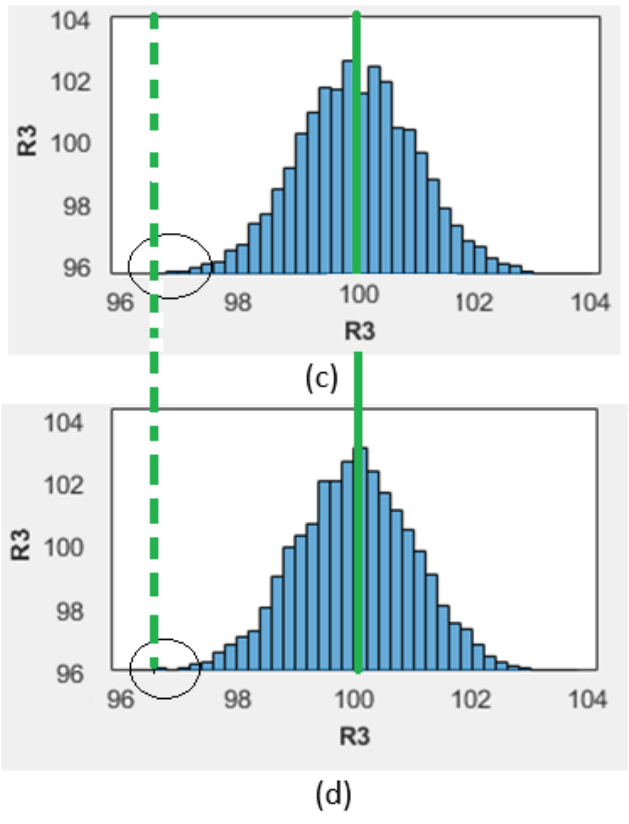


Fig. 8. Probabilities of Occurrences of Changes of Various Resistive Values of R2 (a) without Current Mirror Circuit (b) with Current Mirror Circuit.

VII. Validation of Monte-Carlo Simulation through Simulink

For validation, simulation conditions have been initiated and designed by using Simulink. Nine strings of red LEDs are tested, in which each of seven LED strings is equivalent to 95Ω, while 96Ω and 100Ω for another two LED strings, respectively. The 300 kHz of PWM DC-DC power circuit is employed to supply the current source to LED modules. Three configuration circuits have been used for comparison. Whereas the first configuration is the string with direct dimming circuit, the second configuration is the string with improvement transistor circuit of CM, and the last configuration is the whole proposed CM circuit (association improvement of transistor circuit and super-diode) as shown in Fig. 9, 10 and 11, respectively. The load currents through the red LEDs strings have been classified as I_n, where n defines the branch number. The measured load currents are compared and discussed accordingly.

From the analysis, it has been observed that the outputs in each individual case consist of two steady states, in between them there exists a transitional period. The first steady-state occurs after the initial current flow across the voltage barrier imposed by the dimming circuit, in each individual string. After the establishment of the first steady-state across each load with reference to its particular load current, the load current passes throughout the transient period to adjust/balance itself, concerning other load currents flowing throughout the other loads of the system. To ease elaboration, this detailed analysis of the simulation has shown only load currents designated with I_1, I_2, and I_3 for the particular strings 1, 2 and 3 of the system. These strings have been tested in three different configuration circuits, i.e., through simple dimming circuits, through the combinational circuits of dimming circuits with Q transistors, and the combination of super-diode mechanism. In Fig. 12, load currents for string (I_1) have been recorded. It has been established that after turning on the LEDs, after 0.01 second along with smaller transient time, the load currents are increased up to 20.99 mA, 20.59 mA and 20.57 mA for first, second and last configuration circuits, respectively. The transient response of Fig. 12(c), when compared to Fig. 12(a), from first to second steady-state shows a little bit faster response, similarly with Fig. 12(b) take takes least transient time to attain final steady state. Circuit operating with dimming circuit only shows smaller values of ripples in their outputs whereas the circuits operating with the combination of Q transistor and dimming circuit take lesser time to reach its steady-state but at the cost of higher values of ripples in their output. Fig. 13(c) shows improvement in the output in terms of lesser transient time lesser ripples in their outputs as compared to the rest of the circuits.

Meanwhile, Fig. 13 depicts the load current responses for I_2. After turning on the system, its load current rises to 22.10 mA, 21.66 mA and 21.63 mA for first, second and last configuration circuits. All load currents transients from first steady-state to second state are comparable for all configuration circuits.

The time responses of I_3 are depicted in Fig. 14, where the load current is increased up to 21.87 mA, 21.44 mA, and 21.41 mA for first, second and last configuration circuits, respectively and showing the comparable load currents transient for all configuration circuits.

As a result, the current flow throughout the load strings is identical and comparable for each string accordingly. By adding a new proposed self-configurable CM circuit, slight improvement has been presented, whereas the range of differences among the strings is computed to see the effectiveness of current sharing in Table I. In the proposed method, the gap differences between minimum and maximum load currents from turning off and on conditions slightly shrink to 1.06 mA when compared to the first configuration circuit. Hence the proposed driver design shows the validity of the phenomenon raised by the Monte-Carlo analysis regarding the proposed circuit of current mirror.

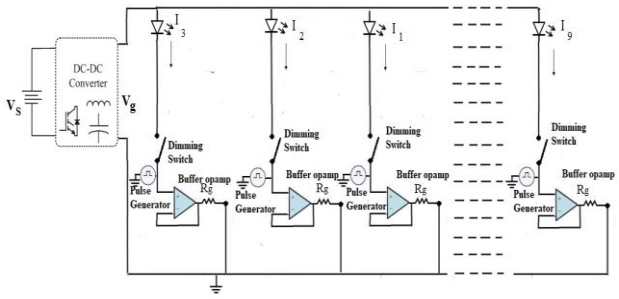


Fig. 9. Dimming Configurable Circuit (First Configuration Circuit).

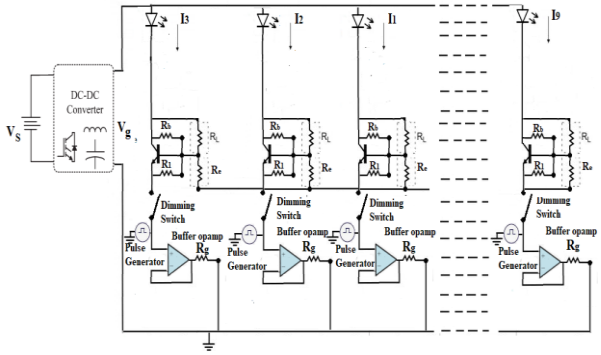


Fig. 10. Dimming Configurable with Improvement Transistor Circuit of CM (Second Configuration Circuit).

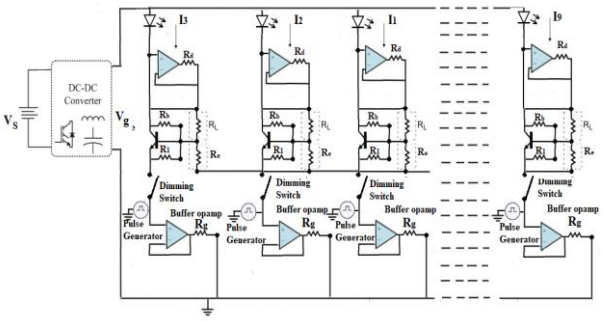


Fig. 11. Dimming Configurable with a New Proposed CM Circuit (Third Configuration Circuit).

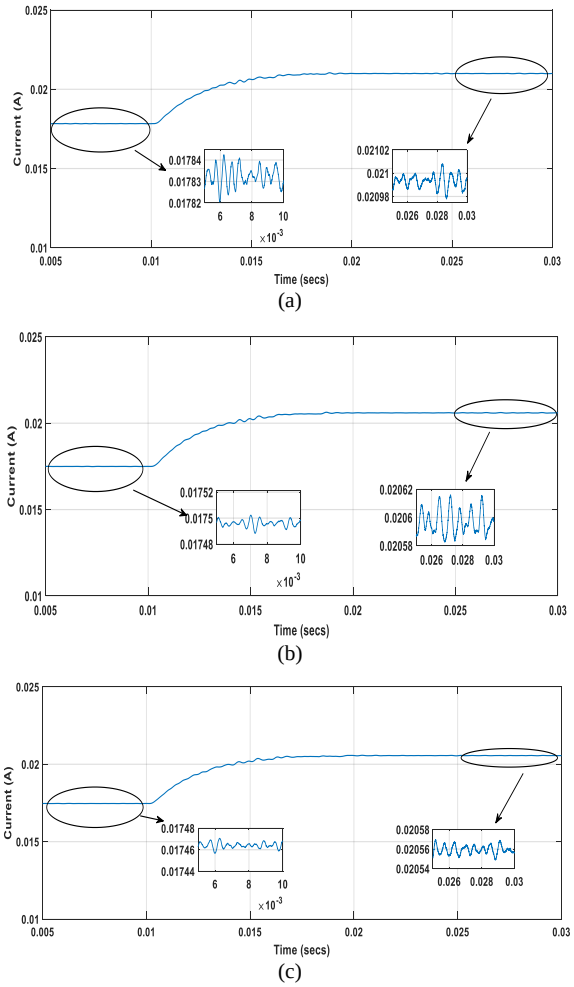
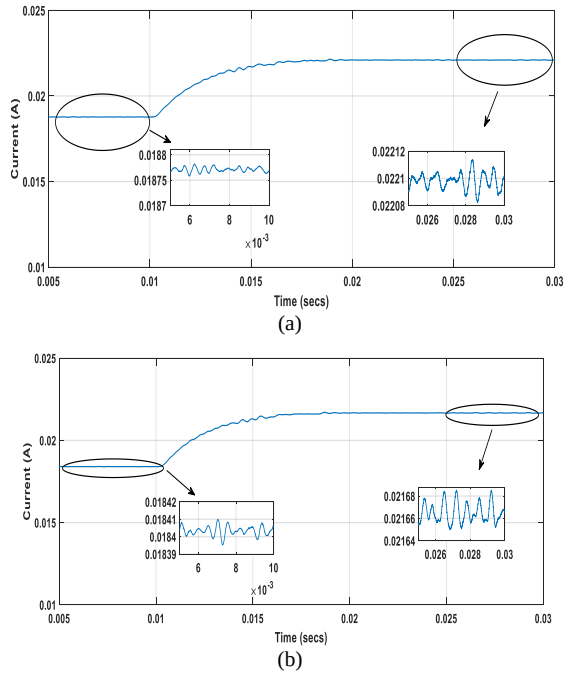


Fig. 12. Load Current Responses (a) First Configuration Circuit, (b) Second Configuration Circuit (c) Last Configuration Circuit.



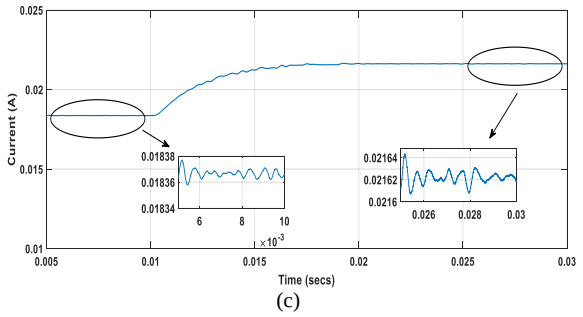


Fig. 13. Load Current Responses (a) First Configuration Circuit (b) Second Configuration Circuit (c) Last Configuration Circuit.

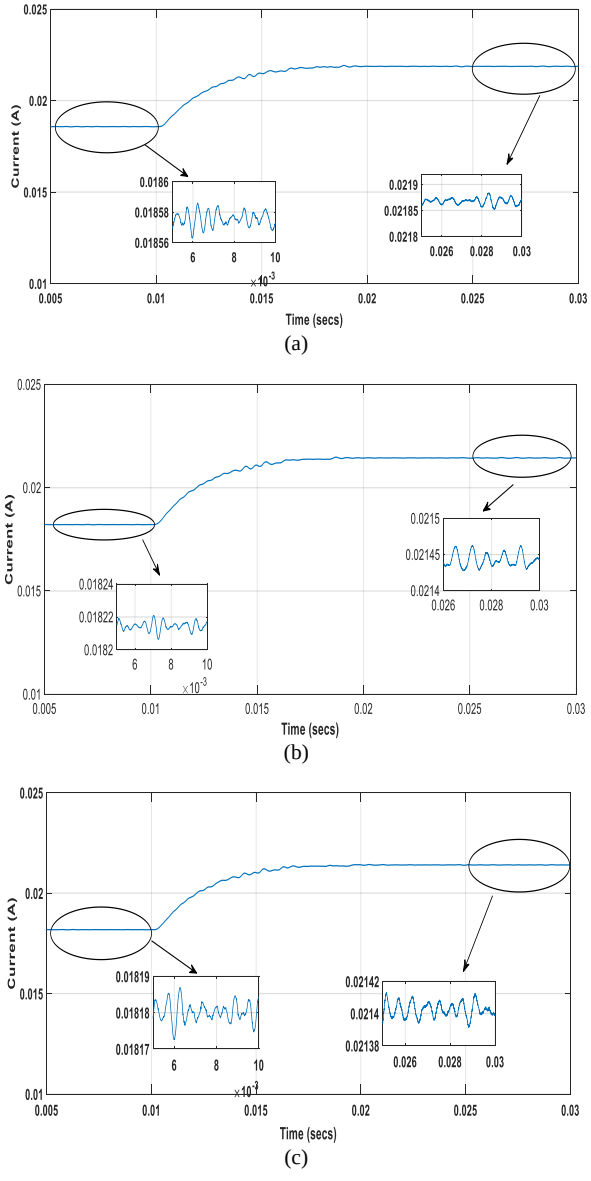


Fig. 14. Load Current Responses (a) First Configuration Circuit (b) Second Configuration Circuit (c) Last Configuration Circuit.

TABLE I. COMPARISON OF THE RANGE DIFFERENCES OF LOAD CURRENTS FOR I1, I2 AND I3

First configuration circuit	Second configuration circuit	Third configuration circuit (a new proposed self-configurable CM circuit)
I_1=20.99 mA	I_1=20.59 mA	I_1=20.57 mA
I_2=22.10 mA	I_2=21.66 mA	I_2=21.63 mA
I_3=21.87 mA	I_3=21.44 mA	I_3=21.41 mA
Maximum current difference =1.11 mA	Maximum current difference =1.07 mA	Maximum current difference =1.06 mA

Conventional current-mirror circuits need a buck converter to trade in with the one constant current load. The second topology to trade in with upgraded self-adjustable current-mirror methods that can address different LED loads under different conditions with the help of one buck converter. The working principle spin around an effective as well as self-configurable merged circuit of transistor and op-amp based current-balancing circuit, along with their dimming circuits. The suggested circuit assures uniformity at the circuit's outputs. This particular scheme of current-balancing circuits excluded the requirement for distinct power supply to regulate the load currents through different kinds of LEDs, i.e., RGB LEDs. The proposed methods are identical and modular, going up to any number of connected corresponding current sources. The methodology has been proficiently examined in the environment of Simulink to substantiate the current balancing phenomenon in parallel LED strings.

VIII. EXPLORATORY ANALYSIS

In color LED driver setups, two different and separate supply voltages, 2.3 V and 3.9 V are supplied to the red LED strings and blue/green LEDs, respectively while using a converter (based on dc-dc conversion). The transients of load currents are observed. The load currents for red LEDs and blue/green are configured from 180 mA to 90 mA to see the transient response. The I_1 as well as I_2 are captured accordingly as depicted in Fig. 15. With the help of the results, it has been noticed that by engaging the proposed self-configuration CM circuit, the transient current responses are slightly fast and comparable with [2].

Regarding equation (6), the losses are computed based on the load current and transistor as follows. In which, 45.0 mW dissipated across the red LED (P_{LED}) and 1.10 mW dissipated across the transistor. Therefore, η is obtained up to 97%. Similar computation could be done for blue and green LEDs.

From the results, it has been observed by employing the proposed self-configuration CM circuit with dimming circuit, the transient responses are faster than the work discussed by Jabbar Hasan in 2011in Table II.

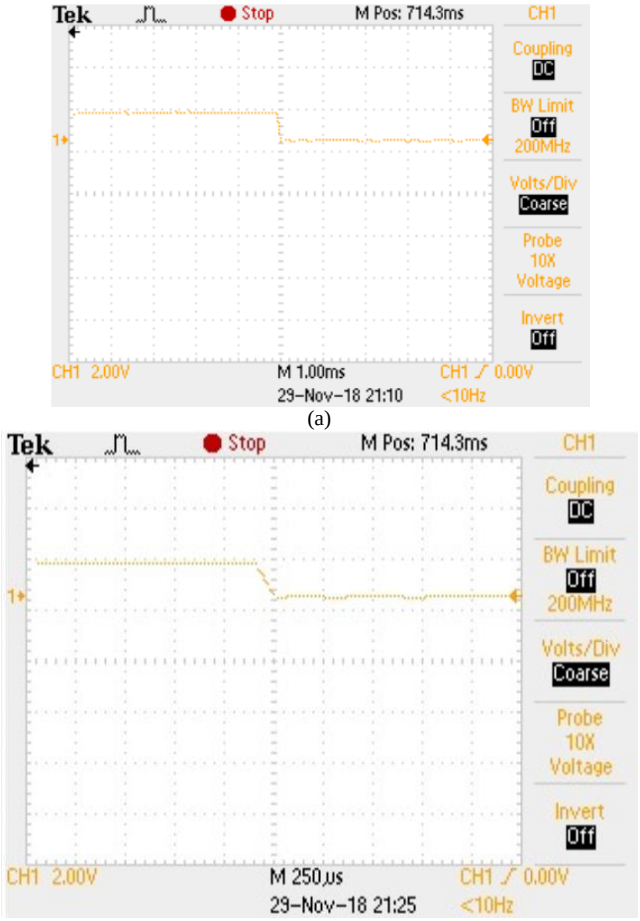


Fig. 15. (a) Transient of the Load Current of I_1 (Load Current for Red Led) from 180 mA to 90 mA (b) Transient of the Load Current of I_2 (Load Current for Green/Blue) from 180 mA to 90 mA.

TABLE II. COMPARISON WITH THE PREVIOUS WORK (J. HASAN, 2012)

	Total delay	Optimization Mode	Accuracy of desired load current (m Ampere)	Number of counts	Complex design
Jabar Hassan	25 ms for red LED Not available for green/blue LED	available	poor	9	yes
Our proposed method	50 µs for red LED 250 µs for green/blue LED	not needed	better accuracy	8	no

IX. Dimming with Efficiency Option

It has been observed that the proposed circuit has been validated to hold the minimization issue of the current imbalances between the loads but lacks for addressing the dimming with good efficiency. To accommodate the issue of efficiency, a tradeoff is needed between the level of accuracy versus efficiency, for high accuracy, there is no need to do any modification but in case of getting high efficiency a modification from the circuit in Fig. 16. is implemented, which shown in the following Fig. 16 by placing R_a (10 ohms), in between the designated node 1 and node 2. By placing R_a, the base Q transistor becomes more active in passing the load current through it and developing lesser voltage at its terminals. Furthermore, it has been noticed that such a combination gives 99% efficiencies at lower dimming values.

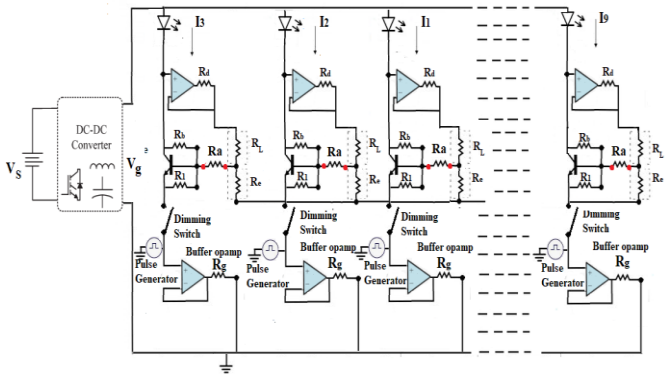


Fig. 16. Addition of $\boldsymbol{R_a}$ in the Proposed CM Circuit.

When increasing dimming (reducing current through the load), more voltage is passed through the pulse generator to the emitter of the transistor but at the same time, more voltage is also applied to the base hence power losses occur across the transistor but while modification Ra is placed at the base in series, then during the process of dimming, there is no increase in base current and consequently low power losses across the transistor terminals. When two different resistive values i.e., 90 and 100 ohms are used, then the maximum difference in their currents comes around 2 mA, which is found higher as compare to Table I.

X. Conclusion

The channel color LEDs driver system with precise current balancing has been verified for red color–green color–blue color-based light-emitting-diode (LED) system. The suggested driver has been checked in order to keep the voltage drop minimum over the LEDs and its associated transistor (current controllers) to keep it in regulation by letting less consumption of power across the transistor, leading to reduced power dissipation in the whole LED's string and increased efficiency in the LED's string. The suggested LED driver system effectively dim individual LEDs in the driver through the individual controller. Calculated efficiencies are 97% and 98.55% for the red color and green/ blue color LEDs, respectively, at their maximum rated currents of 21.0 mA and 22.0 mA. Two different drive voltages for the driver and the current control in the individual LEDs are used, leading to an increase in the strength of the driver which consisting of multiple LEDs. Furthermore, it has been observed that the configuration of super diode along with its associated CM circuits are better in terms of time response. Lastly, the counts in this circuit have been greatly reduced as compared to its predecessors. The proposed circuit facilitates the user to do dimming at the string level and in the whole set of LEDs. It has been noted that the effectiveness of current sharing can also be applied for red or green or for blue color of LEDs.

Conflict of interest: The authors declare no conflicts of interest.

References

[1] Shaheer Shaheer Shaida Durrani, Abu Zaharin Ahmad: An efficient Digitally Controlled for RGB LED Driver, 2017 DOI:10.1109/ICETAS.2017.8277843 Conference: 2017.

[2] Jaber Hasan, Simon S. Ang:A High-Efficiency Digitally Controlled RGB Driver for LED Pixels" IEEE Transactions on Industry Applications, VoL. 47, No. 6, November/December, 2011.

[3] H. van der Broeck, G, Sauerlander, M. Vendt, Power driver topologies and control schemes for LEDs, in Proc. IEEE APEC, 2007.

[4] M. Doshi, R. Zane, Control of solid-state lamps using a multiphase pulse width modulation technique, IEEE Trans. Power Electron., vol. 25, no. 7, pp. 1894–1904, Jul. 2010.

[5] Sung-Jin Choi, Adaptive Current-Mirror LED Driver employing Super-diode Configuration, 2014 IEEE International Conference on Industrial Technology,2014.

[6] Yen-Chung Huang, Hung-Wei Chen: A Novel Fast-Switching Current-Pulse Driver for LED Backlight Applications" 2016 International Conference on Consumer Electronics-Taiwan,2016.

[7] S.-J. Choi, T.-H. Kim: Symmetric Current- Balancing Circuit for LED Backlight with Dimming, IEEE Trans. Industrial Electronics, Vol. 59, No. 4, pp.1698-1707,2012.

[8] Kang Hyun Yi School of Electronic and Electric Engineering, Daegu University "High Voltage, Low Current High-Power Multichannel LEDs LLC Driver by Stacking Single-Ended Rectifiers with Balancing Capacitors", MDPI, 23 March 2020.

[9] Bhawna Aggarwal, Maneesha Gupta a, A.K. Gupta b a Netaji: A comparative study of various current mirror configurations: Topologies and characteristics, Subhash Institute of Technology, Delhi University, Sector-3, Dwarka, New Delhi 110078, India b National Institute of Technology, Kurukshetra, Haryana, India,2016.

[10] Pedro S. Almeida, Joao M. Jorge, Claudio R.B.S. Rodrigues, Guilherme M. Soares, A Novel Method of Current Equalization in LED Strings Based on Simple Linear Circuit, IEEE International Symposium on Industrial Electronics, June 201 I.

[11] Chen, Poki, Yung-Hsuan Chen, John Carl Joel S. Marquez, Ruei-Ting Wang, Jiann-Jong Chen, and Yuh-Shyan Hwang. "Low Flicker Dimmable Multichannel LED Driver With Matrix-Style DPWM and Precise Current Matching." IEEE Transactions on Very Large Scale Integration (VLSI) Systems 28, no. 11 (2020): 2233-2242.

[12] Nadershahi, Shahnad. "Current controller for output stage of LED driver circuitry." U.S. Patent 10,891,893, issued January 12, 2021.

[13] Modepalli, Kumar, and Leila Parsa. "N-color scalable LED driver." U.S. Patent 10,757,775, issued August 25, 2020.

Prediction of Sunspots using Fuzzy Logic: A Triangular Membership Function-based Fuzzy C-Means Approach

Muhammad Hamza Azam[1], Mohd Hilmi Hasan[2], Said Jadid Abdul Kadir[3], Saima Hassan[4]
Centre for Research in Data Science, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia[1, 2, 3]
Institute of Computing, Kohat University of Science and Technology, 26000, Kohat, Pakistan[4]

***Abstract*—Fuzzy logic is an algorithm that works on "degree of truth", instead of the conventional crisp logic where the possible answer can be 1 or 0. Fuzzy logic resembles human thinking as it considers all the possible outcomes between 1 and 0 and it tries to reflect reality. Generation of membership functions is the key factor of fuzzy logic. An approach for generating fuzzy gaussian and triangular membership function using fuzzy c-means is considered in this research. The problem related to sunspot prediction is considered and its accuracy is calculated. It is evident from the results that the proposed technique of generating membership functions using fuzzy c-means can be adopted for predicting sunspots.**

Keywords*—Fuzzy logics; fuzzy c-means (FCM); Gaussian membership function; prediction; sunspots; triangular membership function

I. Introduction

For humans, many tasks are straightforward, like carrying delicate items, passing through crowds or parking a car. Whereas, these tasks can be challenging for computers or machines. Such tasks are made easy and simple for us due to our ability to deal with ambiguous and imprecise data. Therefore, in order to replicate this human operator's control behavior, we must model our systems in a way that make it capable of handling ambiguous and imprecise knowledge. This is exactly what fuzzy logic system do [1], it succeeds where system is overly challenging and complicated. Fuzzy logic has been effectively used in numerous applications from speech [2] and script [3] recognitions to control systems like speed control of Non-Silent Permanent Magnet Synchronous Motors [4].

The word fuzzy applies to details that are ambiguous and not obvious.[5] We sometimes face a condition in real world where we cannot decide if the state is falsifiable, there fuzzy logic offers a very precious reasoning versatility. In this manner, we can identify ambiguities and impreciseness of any scenarios. 1 and o depicts complete truth and false values in Boolean systems. Whereas, there is not completely true or false logic in fuzzy logics. However, in fuzzy logic there are so much of alternating values that are partly true and partly false [6].

One of the most important elements of fuzzy logic system (FLS) is membership function. A membership function is considered as a curve that maps each point value of input space to a membership degree. Input space consist of all the possible elements of consideration in each application, that is also known as the universal set (U) or discourse of universe. Selection of membership function is very critical in fuzzy logic as the entire fuzzy inference system (FIS) depends on the type of membership function used. There are various types of membership function but the most widely used ones are Gaussian, Triangular and Trapezoidal Membership Functions. Hence the generation of membership functions play a vital role in fuzzy logic system.

Few methods are proposed in [7] and [8] to generate single type of membership function. Whereas generation of multiple membership function provides a powerful toolbox for users to solve the problems in more effective ways[9]. Generation of multiple membership function using Complementary-Metal-Oxide-Semiconductor (C.M.O.S) is proposed in [9], which focus on using electric-current for generating membership functions. Based on the knowledge of the author, the literature has not yet recorded ways of generating multiple membership functions by fuzzy c-mean or data driven approach. Therefore, an approach is proposed in this research for generating multiple type of membership functions.

In this paper, we critically analyzed the proposed methods and algorithms in the generation of triangular and gaussian membership functions using fuzzy c-means. The contribution of this paper are as follow: (1) providing practitioners and researchers with insight and future direction on membership function generation through fuzzy c-means, (2) in terms of membership function generation, we investigate the approach of membership function generation using fuzzy c-means.

The remaining paper is arranged is following manner: background of crisp and fuzzy set as well as different current approaches to generate membership function is explained in Section 2, proposed approach and methodology are presented in Section 3. Whereas, Section 4 presents experiment and simulation results and Section 5 present future direction of research and its conclusion.

II. Background

A. Crisp Set

Crisp sets are the basis of classical logic, in which a series of different entities known as a collection. As an example, if we consider colours like blue and black, they are both

different entities considering their characteristic but can be considered as a collection by following the notation {black, blue}. Capital letters are usually used to represent crisp sets, so the earlier example can be represented as:

A = {black, blue}

A crisp subset could be determined from a comprehensive set that contains the values of subset based on certain conditions. Let suppose as an example, we have a set X, such that it contains the integers that are less than 10 and greater than or equal to 2. This set can be represented by the following notation:

X = {a | a is a whole number and $2 \leq a \leq 10$}

The subset represented above can be presented in graphical form if a characteristic notation or indicator function is introduced, which in above case is the operation described over the set of whole numbers, that can be represented as A, which shows the membership of values in subset X of A. To attain this, we call the elements of A in X as 1 and others as 0. Hence, the deduced indicator function can be:

$$1_X(a) = \begin{cases} 1 \; if \; 2 \leq a \leq 10 \\ 0, otherwise \end{cases} \quad (1)$$

It can be illustrated as in Fig. 1.

B. Fuzzy Set and System

In 1965 Lofti Zadeh presented the concept of fuzzy sets, which led to the foundation of fuzzy logics and system [10]. In fuzzy logic theory, a fuzzy set is a set that ensures partial membership of a set. which can be determined by the degree of membership (μ)[11], which represent all the values between 1(a value fully belong to set) and 0 (a value totally not belong to set). It's obvious that if we eliminate all the belonging values except for 1 and 0, fuzzy set will be converted and act as a crisp that which was defined earlier.

The membership functions (MF) of a set is a link between the components of set and their belong to degree [12]. Fig. 2 shows the graphical representation of fuzzy membership function over temperature representation.

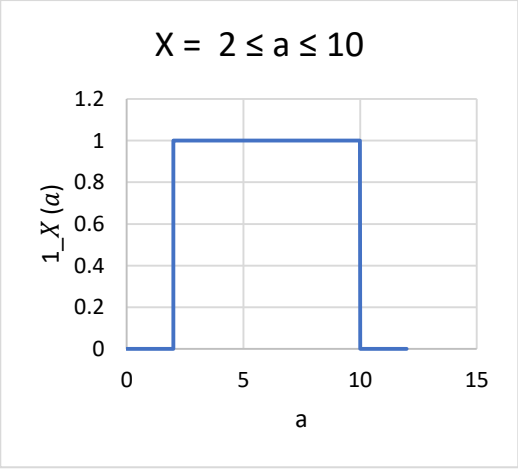


Fig. 1. Crisp Data Representation.

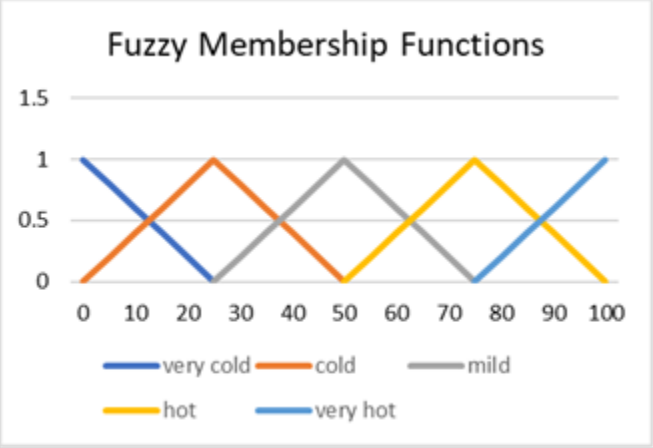


Fig. 2. Fuzzy Data Representation.

Fuzzy system for the temperature is illustrated in Fig. 2 that depicts how a human feel the temperature in real life that can be divided into sub-categories as "very hot", "hot", "mild", "cold" and "very cold", where as if we consider conventional crisp or Boolean logic we have only two possibilities that are hot or cold. Here we can see that at 40 degree of temperature, the system could be defined as being cold to a factor of 0.4 and mild to a factor of 0.6.

C. Fuzzy Inference System

A fuzzy framework is a pool of fuzzy expert intelligence and knowledge base that rather than the conventional and definite Boolean logic, can trigger data in vague terms. Expert knowledge is a combination of both fuzzy membership functions (MF) and fuzzy rule-base that contains all the fuzzy rules and are of form: If (conditions are true or fulfilled) Then (consequences are inferred) [1].

The basic architecture of fuzzy system is presented in Fig. 3.

A fuzzy system consists of 4 main components, namely a fuzzifier, an inference engine, fuzzy knowledge base and a defuzzifier.

1) Fuzzification: Fuzzification is the prime step of fuzzy system where a crisp input is converted to fuzzy input [13] and is assigned a membership degree. The effective working of fuzzy logic system depends on the efficiency of membership function generated in fuzzification process.

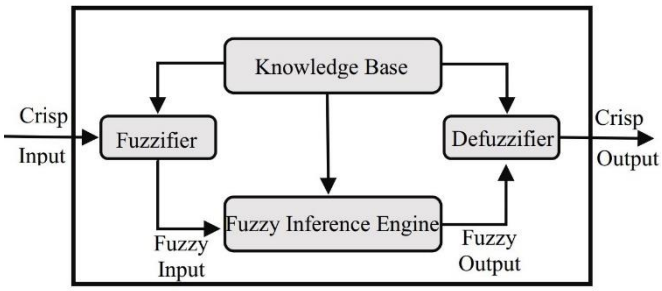


Fig. 3. Fuzzy Inference System (FIS).

2) Fuzzy-Inference-Engine: It matches fuzzy input degree based on fuzzy rule and make decisions related to the application on rules to input field. The inference engine act as a mind of fuzzy system as it provides the decision-making logics of the system, it can be implied as an emulation of human thinking and decision making.

3) Knowledgebase: It contains IF-THEN statements and fuzzy rules that helps in decision making based on linguistic information.

4) Defuzzification: It is the final stage of fuzzy system where a fuzzy set obtained by inference engine is converted to crisp output [1].

D. Fuzzy Membership Functions and its Types

Membership function can be described as a curve which determines how membership values between 1 and 0 are mapped to input space usually called universal set, that consist of all possible values of interest in an application [11]. There are 3 major and widely used types of membership functions.

1) Fuzzy triangular membership function: Fuzzy triangular MF can be described using {a, b, c} as:

$$\mu_F(x;a,b,c) = \begin{cases} 0; x \le a \\ \frac{x-a}{b-a}; a < x \le b \\ \frac{c-x}{c-b}; b < x < c \\ 0; x \ge c \end{cases} \quad (2)$$

Another representation based on min and max is as follow:

$$\mu_F(x;a,b,c) = max\left(min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (3)$$

Here a < b < c represents the coordinates of triangular MF on x-axis. Graphical representation of triangular MF is presented in Fig. 4.

2) Fuzzy trapezoidal membership function: Fuzzy trapezoidal MF can be described using {a, b, c, d} as:

$$\mu_F(x;a,b,c,d) = \begin{cases} 0; x \le a \\ \frac{x-a}{b-a}; a < x < b \\ 1; b \le x \le c \\ \frac{d-x}{d-c}; c < x < d \\ 0; x \ge d \end{cases} \quad (4)$$

Another representation based on min and max is as follows:

$$\mu_F(x;a,b,c,d) = max\left(min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right) \quad (5)$$

Here a < b < c < d) represents coordinates of trapezoidal MF on x-axis. Graphical representation of trapezoidal MF is presented in Fig. 5.

3) Fuzzy Gaussian membership function: Fuzzy Gaussian MF can be described using (σ, c) as:

$$f(x;\sigma,c) = e^{\frac{-(x-c)^2}{2\sigma^2}} \quad (6)$$

A cluster center c and width value σ are used to describe Gaussian MF. Graphical representation of Gaussian MF is presented in Fig. 6.

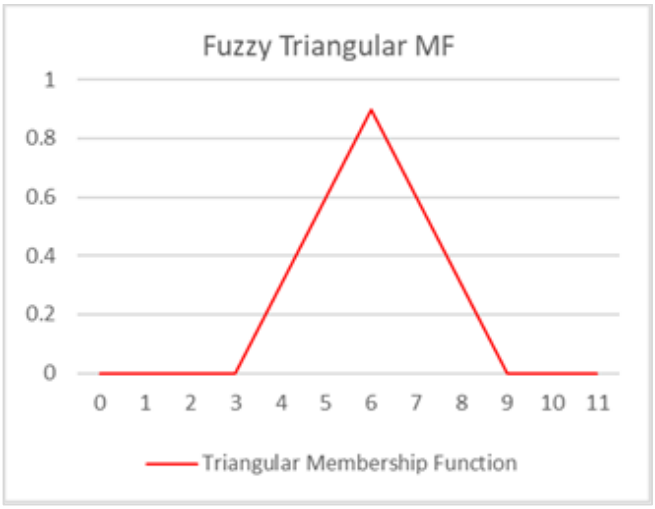


Fig. 4. Triangular Membership Function.

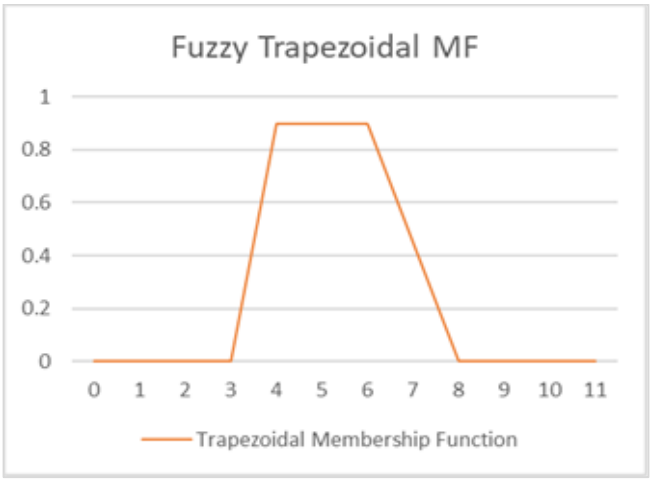


Fig. 5. Trapezoidal Membership Function.

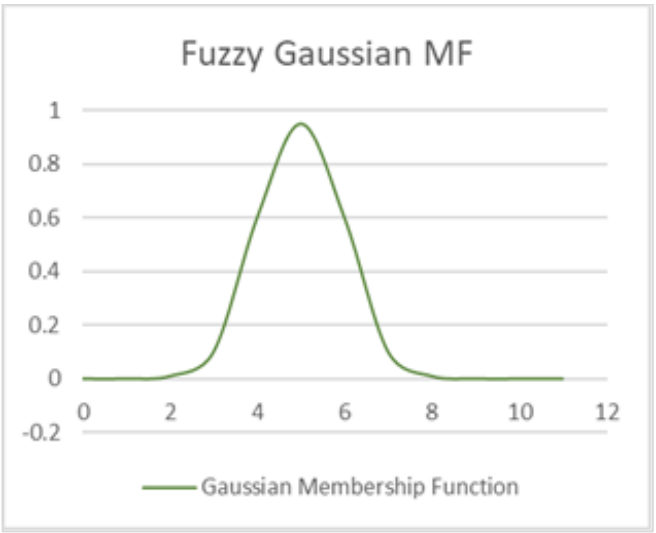


Fig. 6. Gaussian Membership Function.

a) Fuzzy clustering: MF can be developed by mean of two method, one is expert knowledge approach where a

membership function is developed based on the parameters suggested by experts, whereas second method focuses on developing using machine learning approach like classification and clustering. As we know that in expert knowledge approach, parameters are directly provided by experts so that can be incomplete very biased as every expert can suggest different parameters based on their knowledge [14]. Apart from that, it can also be lacking accuracy and expert may have incomplete opinion or may not be available at all time. The method of data processing method will reduce some drawbacks of expert knowledge approach, if not eradicate them. We are therefore concentrating on the generation of fuzzy membership function automatically through data clustering approach in this study.

b) FCM Algorithm: There are many types of clustering algorithms. Fuzzy C-means is one of the popular and essential clustering techniques based on fuzzy logic. The working of FCM is described below:

Fuzzy C-means algorithm can work according to the steps mentioned below:

1) Fix cluster centers (c) i.e. ($2 \leq c \leq n$) and select value for parameter n.

2) Initialize partition matrix (fuzzy membership matrix) Uij.

3) Calculate cluster centers (fuzzy centers) for each step.

4) Update partition matrix (membership matrix)

$$Uik = \sum_{j+1}^{c} [(\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}]^{-1} \quad (7)$$

Here m is a facinis (fuzziness) parameter.

1) Check for convergence.

2) If $\| U(k-1) - U(k) \| \leq E2$ stop else return to step 3. Here E2 is Threshold.

The flow chart for this process is presented in the Fig. 7:

FCM Comparison with other Clustering Algorithms.

Apart from FCM some conventional or hard computing algorithms like K-means, Y-means, etc. are also available. Some advantages of FCM over other clustering algorithms are described below:

- For overlapping dataset, FCM offers relatively better results than k-means.
- In FCM, every datapoint and cluster center is assigned, which results in the possession of multiple cluster centers on data points. Whereas, in K-means algorithm data point may only belong to single cluster center. [15].
- FCM can relatively converge more quicker as compare to K-means but that also increases the computational complexity of FCM. [15].
- In some cases or applications, FCM is more effective, robust and consistent in performance as compare to other clustering algorithms. [16].
- A comparison among Y-means, FCM and K-means is done which shows that FCM produces better results. [16].
- For intrusion detection, the key benefit of FCM is its high detection accuracy and low false positive rate. FCM is an effective approach but also time-consuming [16].

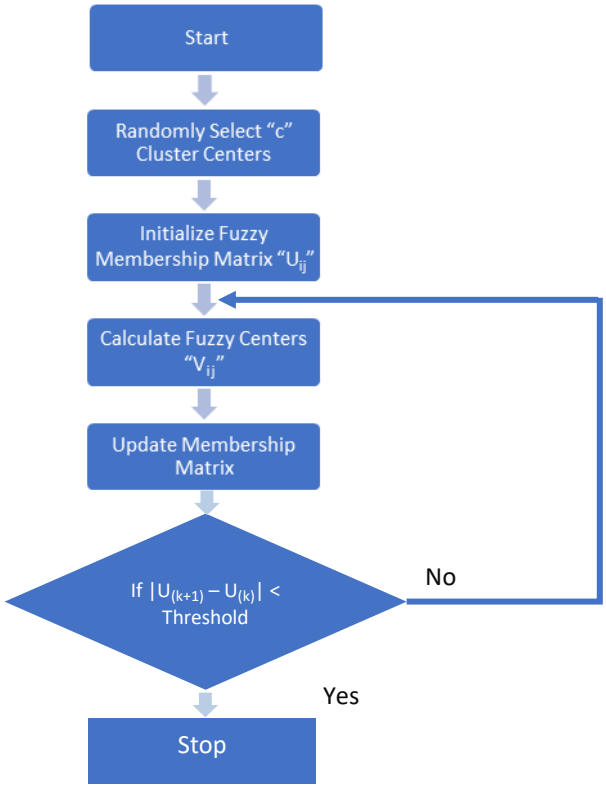


Fig. 7. FCM Flowchart Diagram.

III. Proposed Methodology

The proposed approach of generating triangular membership function is depicted in Fig. 8 and is discussed below:

To generate Gaussian and Triangular membership functions. A dataset is passed through fuzzifier where crisp inputs are converted to fuzzy inputs using FCM. Once data is passed through FCM, it gives fuzzy membership Matrix (U-matrix) and cluster center as an output. Gaussian membership function is than approximated using these values of membership matrix and cluster center. U-matrix and cluster centers are than passed through the formula sets where it converts it into the parameters needed for approximating triangular membership function. Once both gaussian and triangular membership functions are approximated they are than evaluated against test data sets to validate the outcome of an approach. The formula set for generating parameters of triangular membership function is presented below.

$a = \alpha - \beta * \gamma;$

$b = \alpha;$

$c = \alpha + \beta * \gamma;$ (8)

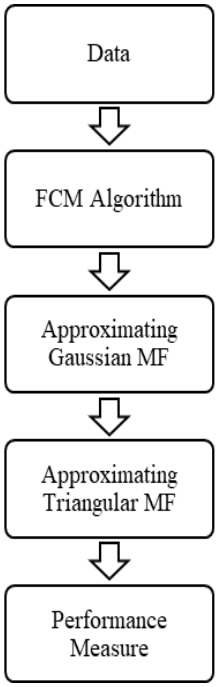


Fig. 8. Proposed Methodology.

Here, a and c represent the floor whereas b represent the peak value of triangular membership function. α and γ are the values calculated from U-matrix and cluster centers while as β is constant.

IV. Results and Discussion

To validate the approach, a prediction data set is used to predict monthly sunspot [17]. Sunspot dataset is used that calculate values of sunspot from 1749/01/01 to 2017/08/31 containing 2820 data values recorded on monthly basis. In the Sun's photosphere, sunspots are transient occurrences which appears as darker spots than the surrounding areas. These regions are of lowered surface temperature that happens due to convection-inhibiting magnetic flux concentrations. It normally occurs in pairs of opposite magnetic polarities. Their numbers vary as per the solar cycle that lasts around 11 years.

FIS is developed for Gaussian membership functions based on sunspots datasets and are presented in Fig. 9.

FIS is developed for Gaussian membership functions based on sunspots datasets and are presented in Fig. 10.

To evaluate the FIS, prediction test on sunspot is performed on sunspot dataset, whose results are shown in Fig. 11:

It can be seen from the results shown in figure above that the prediction results are calculated for sunspot dataset for both Gaussian and triangular membership functions based on training and testing datasets. If we analyze the results, Gaussian membership function has the prediction accuracy of 99.46% whereas triangular membership function has an accuracy of 99.50%. It is evident that here Triangular membership function outperform Gaussian membership function and hence produce better results.

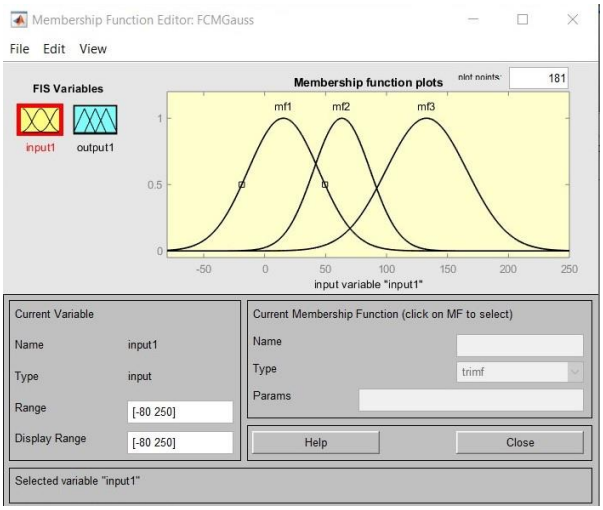


Fig. 9. Fuzzy Gaussian FIS.

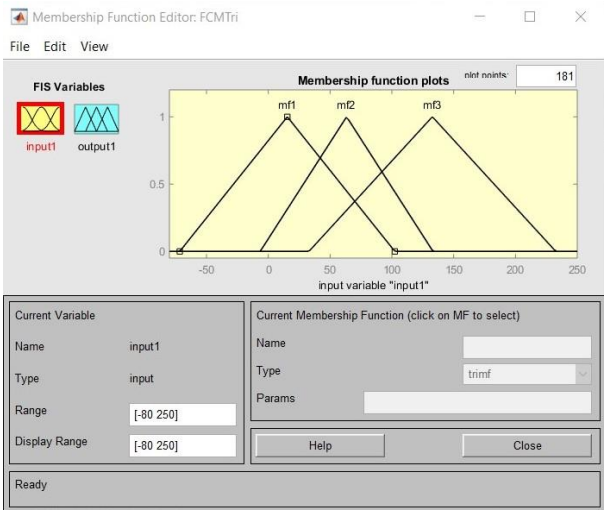


Fig. 10. Fuzzy Triangular FIS.

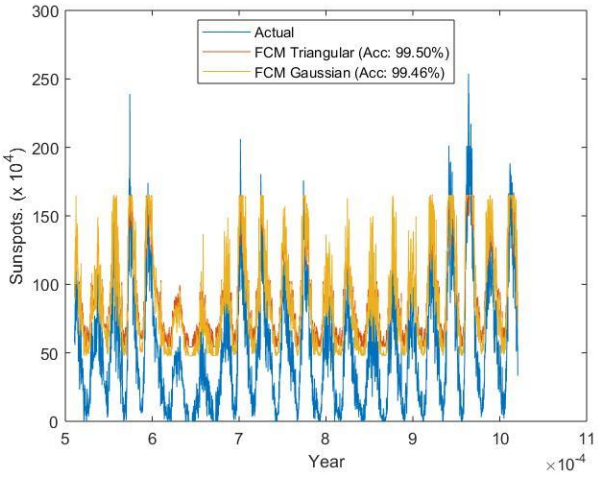


Fig. 11. Sunspots Prediction Results.

V. Conclusion and Future Work

In this research, a technique is proposed to generate triangular and Gaussian fuzzy membership functions through fuzzy c-means. By analyzing the results, it can be concluded that the proposed approach of generating triangular and Gaussian membership functions using fuzzy c-means can be used for prediction of sunspots. This approach will be very effective in the field of data science and specially for prediction problems. It can have its application in many real world's data science problems such as classification, regression, and prediction. The approach will be applied to solve prediction problems such as to forecast electricity demand and price. In future we target to generate triangular as well as trapezoidal membership function using fuzzy type 2 and fuzzy interval type to systems.

Acknowledgment

This research is an ongoing research supported by Fundamental Research Grant Scheme (FRGS/1/2018/ICT02/UTP/02/1); a grant funded by the Ministry of Education, Malaysia.

References

[1] Zuliana and A. M. Abadi, "Sugeno fuzzy inference method and matlab application program for simulation of student performance evaluation in the elementary mathematics learning process," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 4223–4228, Jul. 2020.

[2] A. M. A. Al-Jawadi, "Speech Recognition and Retrieving using Fuzzy Logic System," Tikrit J. Pure Sci., vol. 15, no. 3, 2010.

[3] D. J. Ostrowski and P. Y. K. Cheung, "A Fuzzy Logic Approach to Handwriting Recognition," in Fuzzy Logic, Vieweg+Teubner Verlag, 1996, pp. 299–314.

[4] M. Usama and K. Jaehong, "Simplified Model Predicted Current Control Method for speed control of Non-Silent Permanent Magnet Synchronous Motors," in 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies: Idea to Innovation for Building the Knowledge Economy, iCoMET 2020, 2020.

[5] H. K. Molia and A. D. Kothari, "Fuzzy Logic Systems for Transmission Control Protocol," in Lecture Notes in Electrical Engineering, 2020, vol. 602, pp. 553–565.

[6] M. Ma, T. Wang, Q. Jianbin, and H. R. Karimi, "Adaptive fuzzy decentralized tracking control for large-scale interconnected nonlinear networked control systems," IEEE Trans. Fuzzy Syst., pp. 1–1, Jul. 2020.

[7] R. Khayatzadeh and M. B. Yelten, "A Novel Multiple Membership Function Generator for Fuzzy Logic Systems," in SMACD 2018 - 15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design, 2018, pp. 101–104.

[8] S. Moshfe, A. Khoei, K. Hadidi, and B. Mashoufi, "A fully programmable Nano-Watt analogue CMOS circuit for Gaussian functions," in 2010 International Conference on Electronic Devices, Systems and Applications, ICEDSA 2010 - Proceedings, 2010, pp. 82–87.

[9] B. Mesgarzadeh, "A CMOS implementation of current-mode min-max circuits and a sample fuzzy application," in IEEE International Conference on Fuzzy Systems, 2004, vol. 2, pp. 941–946..

[10] George J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic. 1994.

[11] M. H. Azam, M. H. Hasan, S. Hassan, and S. J. Abdulkadir, "Fuzzy Type-1 Triangular Membership Function Approximation Using Fuzzy C-Means," in 2020 International Conference on Computational Intelligence (ICCI), 2020, pp. 115–120.

[12] M. H. Hasan, J. Jaafar, and M. F. Hassan, "Fuzzy C-Means and two clusters' centers method for generating interval type-2 membership function," in 2016 3rd International Conference on Computer and Information Sciences, ICCOINS 2016 - Proceedings, 2016, pp. 627–632.

[13] K. O. Oluborode, O. O. Obe, O. Olabode, and O. T. Adeboje, "Adaptive neuro-fuzzy controller for double lane traffic intersections," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 6455–6463, Jul. 2020.

[14] S. S. Jamsandekar and R. R. Mudholkar, "Fuzzy Classification System by Self Generated Membership Function Using Clustering Technique," BIJIT - BVICAM's Int. J. Inf. Technol., vol. 6, no. 1, pp. 697–704, 2014.

[15] Jyotismita Goswami, "Inpatient child and adolescent therapy groups: Boundary maintenance and group function," Int. J. Sci. Eng. Appl. Sci., vol. 1, no. 2, pp. 170–178, 2015.

[16] T. Singh and M. M. Mahajan, "Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 4, no. 5, p. 2277, 2014.

[17] "Sunspots | Kaggle." [Online]. Available: https://www.kaggle.com/robervalt/sunspots. [Accessed: 30-Nov-2020].

Optimum Spatial Resolution of Satellite-based Optical Sensors for Maximizing Classification Performance

Kohei Arai
Faculty of Science and Engineering
Saga University, Saga City
Japan

***Abstract*—Optimum spatial resolution of satellite based optical sensors for maximizing classification performance is investigated. Also, classification performance assessment method considering spatial resolution of satellite based optical imagers is proposed. Optimum spatial resolution which makes the highest classification accuracy is determined from spatial frequency components, spectral features of objects and classification method. First, in this paper, based on the relationship between variance of pixels and classification accuracy, classification accuracy for Landsat Multiple Spectral Scanner: MSS images with various Instantaneous Field of View (IFOV) will be shown. In their connection, variance of pixel values for images with various IFOV will be clarified. Second, assuming the shape of boundary line between adjacent categories is circle, relationship among IFOV, ratio of Mixels and classification accuracy will be cleared under the supposition that the number of Mixels equals to that of misclassified pixels. Finally, it will be also shown that aforementioned relationships and optimum spatial resolution have been confirmed by using airborne based MSS data of Sayama district in Japan.**

Keywords*—Spectral information; spatial information; maximum likelihood decision rule; satellite image; image classification; mixed pixel (Mixels); optimum spatial resolution; classification performance; spatial and spectral features

I. Introduction

Optimum spatial resolution of satellite based optical sensors for maximizing classification performance is investigated. From the point of view of classification performance, there must exists an optimum spatial resolution of the spaceborne onboard optical sensors because the variances of the class categories are getting large in accordance with the spatial resolution. The classification performance is getting down because the overlapped areas among the class categories are getting large which results in confusion probabilities are getting large in accordance with spatial resolution.

Since variance of pixels correspond to that in the feature space increases in accordance with improvement of spatial resolution, classification accuracy will be gotten worse in accordance with improvement of spatial resolution under the limitations of variety of objects and class categories. On the other hand, classification accuracy gets better in accordance with improvement of spatial resolution because of decreasing of a ratio of "Mixels" which are pixels composing with plural class categories. Since aforementioned two effects contribute to classification accuracy multiplicatively, it seems that there exists an optimum spatial resolution.

The Instantaneous Field of View: IFOV of the multispectral radiometer mounted on the earth observation satellite with the highest classification accuracy, that is, the optimal spatial resolution, is generally determined by the spatial frequency component, spectral characteristics, classification method, and the like of the observation target [1]-[9]. Classification accuracy is defined as the discrimination efficiency (diagonal element of the confusion matrix) in maximum likelihood classification, and when parameters such as the object to be observed and the number of classes are limited, increasing the spatial resolution generally increases the variance in the feature space [10].

The classification accuracy becomes worse. On the other hand, the ratio of Mixels (mixed pixels) of different classes becomes smaller as the spatial resolution is improved, so that the classification accuracy is improved [11],[12]. Since the above effects synergistically contribute to the classification accuracy, it is considered that there is an optimal spatial resolution.

The motivation of this research study is to clarify relations between spatial resolution of optical sensors onboard satellites and classification performance and then find out optimum spatial resolution for maximizing classification performance.

This paper first clarifies the relationship between the instantaneous visual field and the classification accuracy [13] by deriving the case variance based on the relationship between the variance of pixel values and the classification accuracy. Next, assuming that the shape of the boundary of the same class region is an arc, the relationship between the instantaneous visual field and the ratio of the Mixels is obtained, and further, the relationship between the instantaneous visual field and the classification accuracy is obtained assuming that the ratio of the Mixels is a false recognition rate. Both relationships show that there is an optimal spatial resolution, and this is confirmed by using MSS data of the aircraft that observed Sayama Hills, in Japan.

In the following section, related research works and research background including motivation of the research are

described. Then, the proposed context classification method is described followed by experimental method together with experimental results. After that, concluding remarks and some discussions are described.

II. Related Research Works

Classification by re-estimating statistical parameters based on auto-regressive model is proposed for purification of training samples. [14]. Meanwhile, multi-temporal texture analysis in Landsat Thematic Mapper: TM classification is proposed for high spatial resolution of optical sensor images [15]. On the other hand, Maximum Likelihood (MLH) TM classification taking into account pixel-to-pixel correlation is proposed [16].

Supervised TM classification with a purification of training samples is proposed [17] together with TM classification using local spectral variability is proposed [18]. A classification method with spatial spectral variability is also proposed [19] together with TM classification using local spectral variability [20].

Application of inversion theory for image analysis and classification is proposed [21]. Meanwhile, polarimetric SAR image classification with maximum curvature of the trajectory in eigen space domain on the polarization signature is proposed [22]. On the other hand, A hybrid supervised classification method for multi-dimensional images using color and textural features is proposed [23].

Polarimetric SAR image classification with high frequency component derived from wavelet multi resolution analysis: MRA is proposed [24]. Comparative study of polarimetric SAR classification methods including proposed method with maximum curvature of trajectory of backscattering cross section in ellipticity and orientation angle space is conducted and well reported [25].

Comparative study on discrimination methods for identifying dangerous red tide species based on wavelet utilized classification methods is conducted [26]. On the other hand, multi spectral image classification method with selection of independent spectral features through correlation analysis is proposed [27]. Image retrieval and classification method based on Euclidian distance between normalized features including wavelet descriptor is proposed [28].

Gender classification method based on gait energy motion derived from silhouettes through wavelet analysis of human gait moving pictures is proposed [29]. Also, human gait skeleton model acquired with single side video camera and its application and implementation for gender classification is proposed [30] together with human gait skeleton model acquired with single side video camera and its application and implementation for gender classification [31]. On the other hand, gender classification method based on gait energy motion derived from silhouette through wavelet analysis of human gait moving pictures is proposed [32] together with human gait gender classification using 3D discrete wavelet transformation feature extraction [33].

Image classification considering probability density function based on simplified beta distribution is proposed [34]. Meanwhile, Maximum Likelihood Classification: MLH based on classified result of boundary mixed pixels for high spatial resolution of satellite images is proposed [35]. On the other hand, context classification based on mixing ratio estimation by means of inversion theory is proposed [36].

III. Research Background

A. Relationship between Instantaneous Visual Field and Variance of Pixel Values

According to Friedman et al. [13], the classification accuracy P is determined by the variance σ^2 of the pixel values and the size α in the feature space of each class, which is determined by the classification method.

$$p = 10^{-0.4/\beta}, \beta = \alpha/\sigma \quad (1)$$

where, α is defined as a range of a discrimination threshold in an arbitrary dimension of an arbitrary class (this threshold is determined by a classification method that regards the range of this threshold as the same class). The unit is the digital count value, which is the same as the unit of the standard deviation (σ) of the pixel value. σ^2 changes depending on the instantaneous field of view. If the time series x (t) of the original pixel values is now sampled at the time interval l and the obtained series is g (ξ), σ^2 (a) is shown in Appendix 1. Equation (2) can be expressed as follows.

$$\sigma^2(a) = 2\{\frac{1}{2!}\frac{1}{l}\int_0^l (x(t) - g(\xi))^2 d\xi - \frac{a^2}{4!}\frac{1}{l}\int_0^l x'(t)^2 d\xi + \ldots - \ldots\} \quad (2)$$

That is, it can be seen that σ^2 (a) can be represented by only the even-order terms of a, and can be represented by an equation that is negative from the second term. Therefore, when x (t) is a continuous function, it can be seen that σ^2 (a) becomes a constant value when a approaches o, and decreases at first as an objective line as it increases. Since σ^2 (a) in the range of the instantaneous visual field near the optimal spatial resolution is considered to decrease parabolically from a constant value, here, Eq. (2) is approximated by the second term.

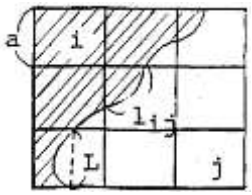
B. Classification Method in Concern

According to Crapper, the Mixel ratio F is expressed by the following equation.

$$F = K_1\sqrt{\pi}(\frac{a^2}{L})(\frac{\sqrt{A}}{A}) \quad (3)$$

where $K_1 = \sum_j L_j/(2\sqrt{\pi}\sum_j \sqrt{A_j}$, L_j denotes the perimeter of the class j area, A_j; the area of the class j area, a; the instantaneous visual field (pixel size), L; the average length when the boundary of the class area that crosses the pixel is approximated by a straight line.

Fig. 1 shows the definition of L. In the figure, the hatched area is defined as class i and the other area is defined as class j. The boundary line length is defined as L_{ij}, and the length connecting the straight lines (broken line length) is defined as L. The average of L is defined as L.



A: pixel size
i, j: class categories
Lij: boundary line length
Lj: approximated straight-line length

Fig. 1. Definition of Notations.

The parameters other than a are obtained by the Crapper [11] as experimental values using 1605 types of sample data as follows. In the process of deriving Eq. (3),

L=0.7935a, K_1=1.82, $\sqrt{A} = 1483(m)$, A=3718600(m^2)

Crapper makes the following assumptions.

1) There are at most two classes in a pixel.

2) The length of the line segment that intersects the edge of the pixel with the boundary between different classes in the image space is not different from the line segment length when it is approximated by a straight-line.

Both of the above assumptions hold if a is sufficiently small compared to the size of the class area. However, when discussing the optimal spatial resolution, it is generally considered that the size of the class area is equivalent to the pixel size. Must be made and this assumption does not hold. In particular, the assumption of (2) often does not hold. Here, it is assumed instead that "the category boundary line in a pixel is an arc." Under this assumption, the ratio q between the average length L_{ij} of the class area boundary line and the average length L obtained by linearly approximating it is the radius r and the pixel size a when the boundary line of the class area is assumed to be circular. It becomes a function and the relationship shown in Fig.2. Considering this ratio and adopting the experimental values of Crapper as other parameters, F can be expressed by the following equation.

$$F = 1.62 * \frac{10^{-3}a}{q\left(\frac{a}{r}\right)}, q\left(\frac{a}{r}\right) = L/L_{ij} \quad (4)$$

Assuming that the classification accuracy of Mixels is 0, the classification accuracy of pixels (pure pixels) composed of a single class is 1, and their combined classification accuracy is equal to (1-F).

C. Overall Classification Accuracy

The total classification accuracy Pt is defined by the following equation in consideration of the influence of the variance of pixel values and the ratio of the Mixels on the classification accuracy.

$$Pt = P\ (1\text{-}F) \quad (5)$$

The pixel size at which this Pt is the highest is defined as the optimal spatial resolution.

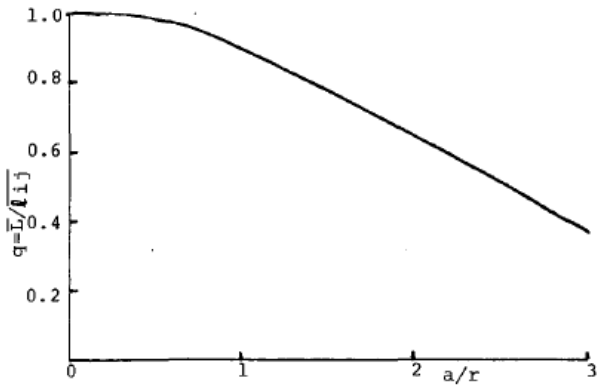


Fig. 2. Ratio of the Line Length of Boundaries between Adjacent Class Categories lij to its Approximated Straight-Line Length L.

IV. Experiment with Aircraft MSS Data Observing Sayama Hills: Example of Optimal Spatial Resolution

A. Geographical Characteristics of the Intensive Study Area

The analysis area is around Sayama as shown in Fig.3. The south side consists of Sayama hills at about 150m above sea level, and the north side consists of flat land about 100m above sea level. The Sayama hill has Sayama lake, broadleaf / coniferous forest, grassland, etc. The flat land consists of development land including bare land, urban area, residential area, paddy field, upland field, tea plantation, etc.

B. Generation and Classification of MSS Data of Instantaneous Field of View: IFOV

On December 12, 1981, data of an instantaneous field of view of 1.25 m was collected using an MSS; DS-1250 for airborne use at an altitude of 500 m. Apply 8 × 8, 12 × 12, 16 × 16, 24 × 4, 48 × 48 pixel window size smoothing filter to this to generate instantaneous visual field data of 10, 15, 20, 30, 60m.

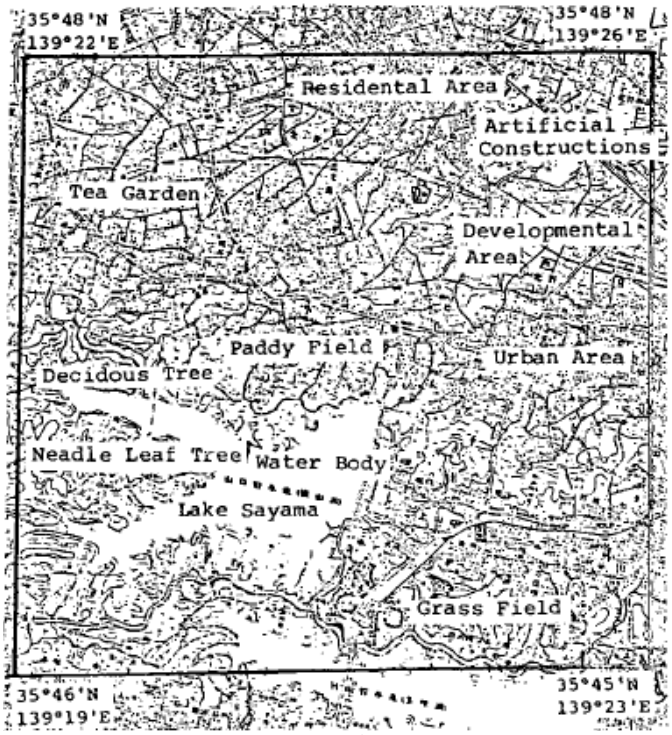
Then, the maximum likelihood classification was tried using each band data in the wavelength range of 0.55 to 0.60, 0.65 to 0.69, 0.8 to 0.89, 8.0 to 14.0 μm, and the discrimination efficiency was evaluated.

Figure 4 shows examples of images of each instantaneous visual field. Table l shows the discrimination efficiency *Pi* for each class by maximum likelihood classification using the data of each instantaneous visual field. *Pi; i* is the class type, the area ratio of each class.

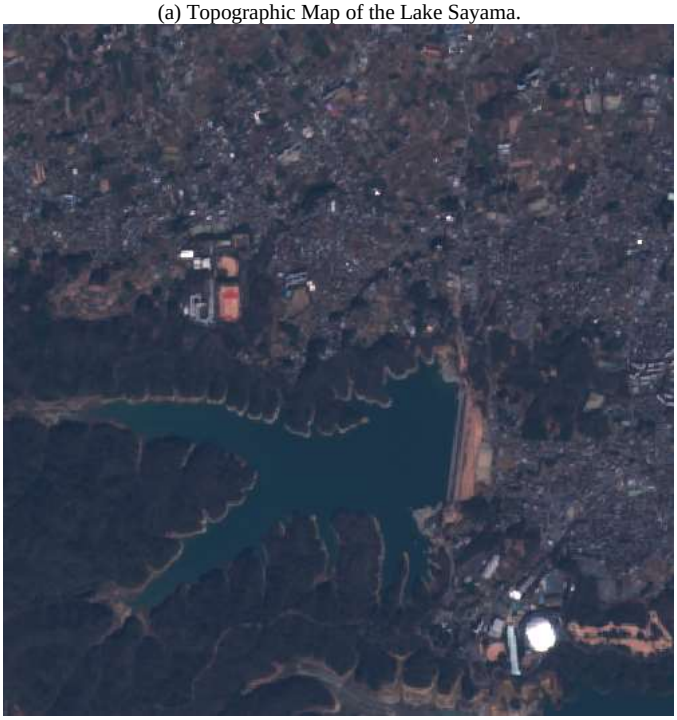
The discrimination efficiency *Pa* at each instantaneous visual field represented by the formula is shown.

$$Pa = \Sigma RiPi \quad (6)$$

Ri is the average of the area ratio of each class obtained by classifying the simulation images of each instantaneous visual field by the maximum likelihood method. From the table, it can be seen that the instantaneous visual field showing the maximum discrimination efficiency differs for each class, and that coniferous forests, fields, development areas, urban areas, etc. have a relatively large number of composite pixels of different classes, and the dimensions of the components are small.

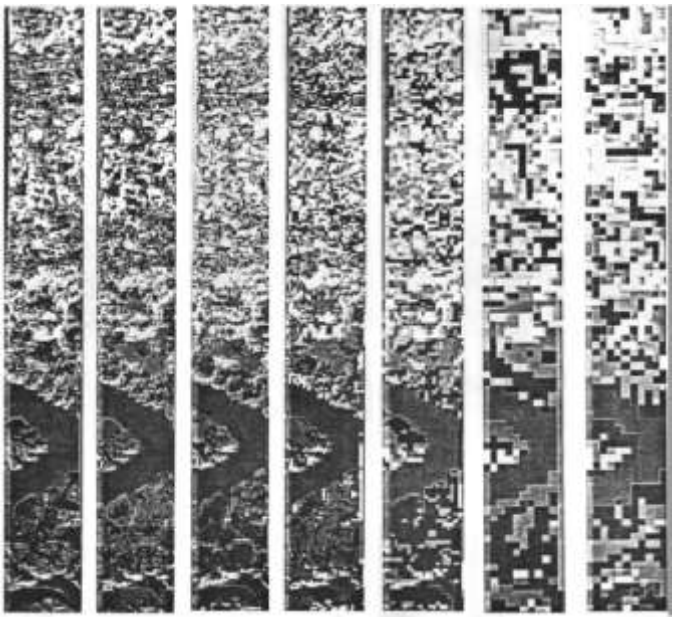


(a) Topographic Map of the Lake Sayama.



(b) Sentinel-2 of the Lake Sayama.

Fig. 3. Location of Study Area and Topographic Feature of the Area.



1.25m Airborne | 10m Simulated | 15m Simulated | 20m Simulated | 30m Simulated | 60m Simulated | 60m Landsat

Fig. 4. Example of Airborne MSS, Simulated MSS with different IFOV and Actual Landsat MSS Image of Sayama, Japan.

TABLE I. CLASSIFICATION ACCURACY FOR THE SIMULATION DATA OF SEVERAL INSTANTANEOUS FIELD OF VIEW (IFOV)

a	1.25	10	15	20	30	60	*Ri*
Needle Leaf Tree	33.1	58.5	87.8	69	80.2	86.2	22.5
Deciduous Tree	76.9	79.4	63.2	91.9	92.4	86.5	17.5
Paddy Field	83.5	84.5	73.4	83	95.5	87	6.4
Perm Area	57.9	91.5	81.8	87	82	45.5	9
Tea Garden	75.6	75.1	88.4	67.9	98.3	93.7	6.9
Grass Field	87	90.9	95.6	80.1	100	88.6	3.2
Developmental Area	38.2	89.5	63.1	98.1	92.8	100	3.6
Water Body	99.5	98.9	99.8	100	99.9	99.8	13.5
Mulberry Field	61.8	86	86.5	92.9	74.6	79.9	7.4
Urban Area	48.8	66.2	88.2	88.6	83.6	70.6	10
Pa	63.7	78.3	83	88.9	88.9	83.5	_____

a: IFOV, i: Class category, *Ri*: Percentage ratio of class category area, *Pa*: Percentage ratio of correct classification

As shown in Fig. 5, it can be seen that the variance of the pixel values is relatively low in the classification accuracy. It was also found that the average classification accuracy in each instantaneous visual field, weighted by the area ratio of each class, was highest when the instantaneous visual field was 30m.

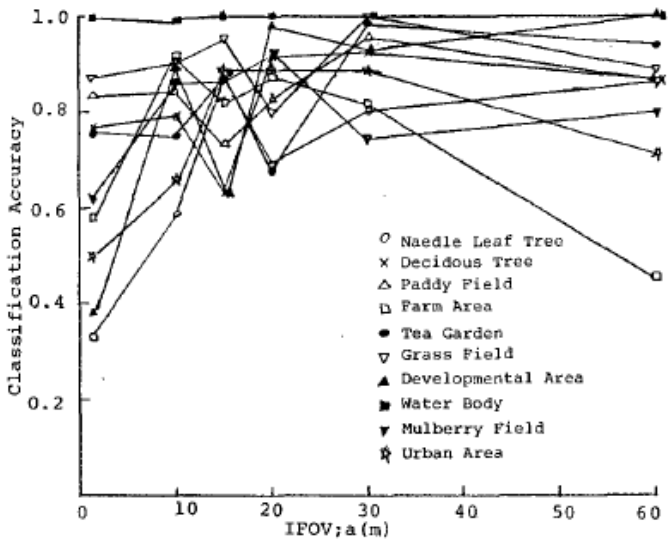


Fig. 5. Classification Accuracies for Various Class Categories.

C. Relationship between IFOV and Variance

As shown in Fig. 6, the variance differs depending on the channel and class; where, in order to clarify the average relationship between the IFOV and the variance, the function was approximated by the least square method.

As a result, Eq. (7) was obtained. The approximation error normalized by the variance value at this time was 0.71.

$$\sigma^2(a) = 498.32 - 0.49a^2 \quad (7)$$

By substituting Eq. (7) into Eq. (1) and evaluating the classification accuracy p using α as a parameter, the result is as shown in Fig. 7. In the figure, the solid line is the calculated value of Eq. (7), and the broken line is the estimated value.

When the aircraft data used in the analysis were classified by the maximum likelihood method, the range α of each class was about 10 to 120. Therefore, Fig. 7 requires the classification accuracy when α is changed in 20 steps from 30 to 90.

D. Relationship between IFOV and Mixing Ratio

Estimated from the land cover classification map of the Sayama area used in the analysis, the average of the radius *r* was determined to be about 20 m when the boundary between each class was assumed to be an arc. Therefore, here, r was changed to 10 to 30 m, and Eq. (4) was opened to evaluate the effect of the change of the Mixels by the instantaneous visual field on the classification accuracy.

Fig. 8 shows the results. The figure also shows the calculated values of Crapper [11] and the experimental values of Jackson [12].

Since the calculated value of Crapper is based on the assumption that the instantaneous visual field is sufficiently smaller than the class size, the calculated value of the model proposed in this paper asymptotically approaches the region where the instantaneous visual field is small. Also, Jackson's experimental values correspond to those where *r* in the calculated values of this model is about 25 m.

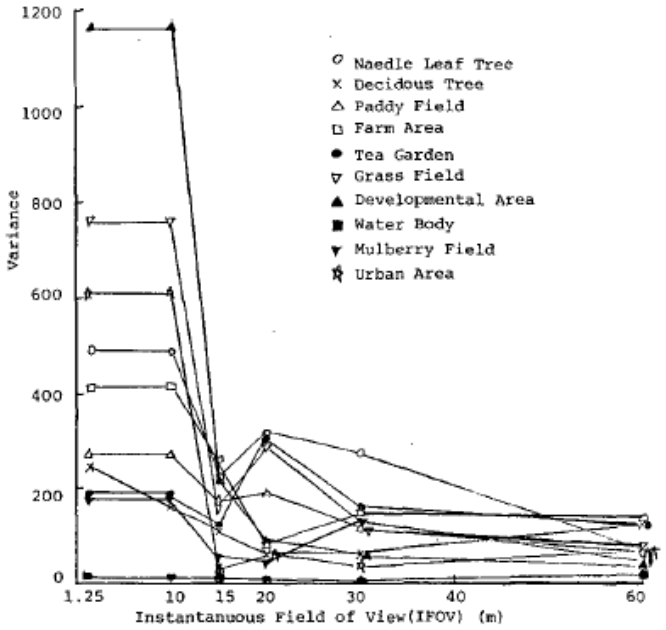


Fig. 6. Variance of Channel 9 Data of each Class Category for each IFOV.

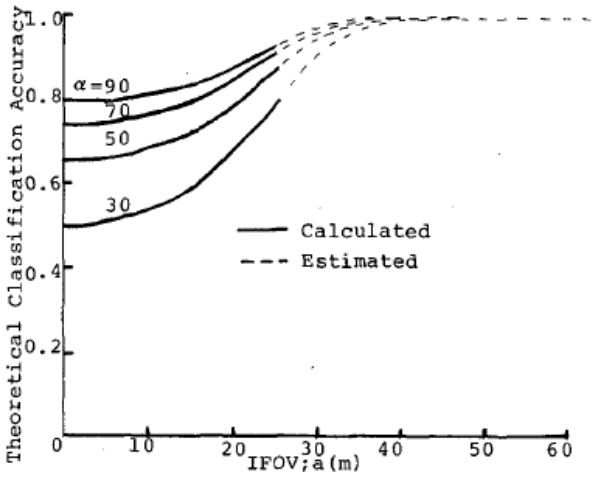


Fig. 7. Classification Accuracy for various Parameter of α.

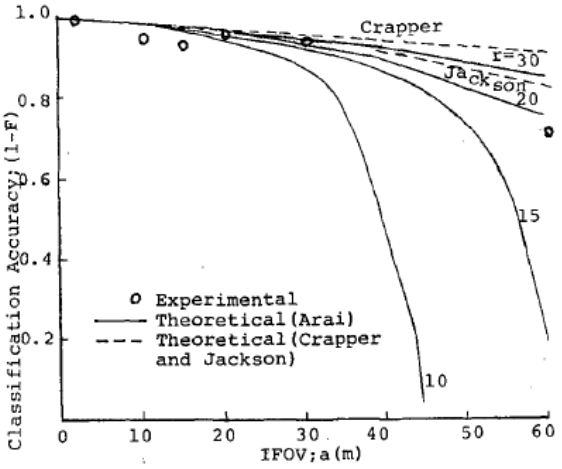


Fig. 8. Decrease of Classification Accuracy in Accordance with Increasing of IFOV.

Comparing the experimental values (circles in Fig. 8) for calculating the ratio of the Mixels for the images of each instantaneous visual field with the calculated values of this model, it can be seen that *r* is about 20 m. This experimental value is based on the assumption that the result of classifying aircraft MSS data with an instantaneous field of view of 1.25 m into 10 classes by maximum likelihood classification is 0 (%) for the Mixels, and this classification result is the MSS image of each instantaneous field of view. In each pixel, the ratio of the number of categories that are composed of a plurality of categories is calculated as the number of Mixels.

This indicates that, when the analysis target area is classified into 10 classes in Table I, the average of the class sizes can be considered to be equivalent to a circle with a semicircle of about 20 m.

E. Relationship between IFOV and Classification Accuracy

The total classification accuracy p_t considering the variance and the Mixels was calculated by Eq. (5) and shown in Fig. 9. At this time, 20m was selected as the parameter r, and 50, 70, and 90 were selected for α. In addition, Fig. 9 also shows the classification accuracy p obtained by actually performing the maximum likelihood classification, and (Table 1). The two values are almost the same, especially when α = 90, which indicates that the method proposed in this paper for finding the optimal spatial resolution is appropriate.

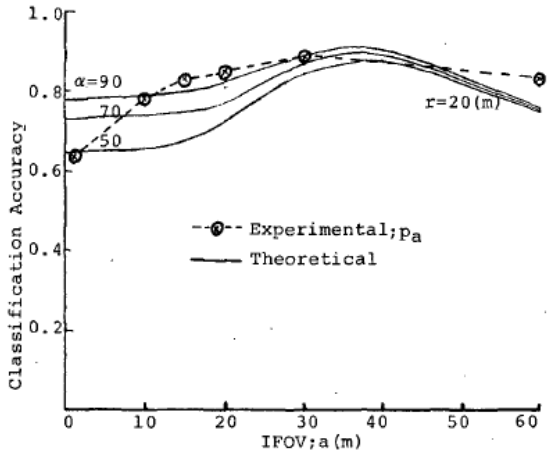


Fig. 9. Comparison of Theoretical Classification Accuracy with Experimental one for Various IFOV.

V. CONCLUSION

Optimum Spatial Resolution of Satellite Based Optical Sensors for Maximizing Classification Performance is investigated. Also, classification performance assessment method considering spatial resolution of satellite based optical imagers is proposed. Optimum spatial resolution which makes the highest classification accuracy is determined from spatial frequency components, spectral features of objects and classification method.

The validity of the method of obtaining the ratio of the Mixels and the variance of the pixel values and the method of deriving the classification accuracy considering them were confirmed by the analysis example of the Sayama area of Japan scene of Landsat satellite imagery data. According to this method, the classification accuracy can be estimated by giving the class size in the feature space and the radius when the class region in the image space is approximated by an arc. It is also possible to find the optimum spatial resolution that maximizes the accuracy.

VI. FUTURE RESEARCH WORKS

The proposed method is adopted in the real earth observation satellite imagery data, and it is a future subject to realize a more usable classification method. Also, optimum spatial resolution would be better to be realized with actual remote sensing satellite based optical sensors.

ACKNOWLEDGMENT

The author would like to express our sincere thanks to Dr. Craper of CSIRO, Prof and Strahler of Hunter College, former Professor Tsuchiya of Chiba University and others for their valuable advices and discussions. The author, also, would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

REFERENCES

[1] Thompson F.J. et al, Final report multispectral scanner data applications evaluation, vol.1, User applications study, ERIM Report No.102800-40I, 1974.

[2] Clark J. et al, Landsat-D thematic mapper simulation using aircraft multispectral scanner data, Proc. 11th ISRSE, ERIM 1977.

[3] Clark J., The effect of resolution in simulated satellite imagery on spectral characteristics and computer assisted land use classification, JPL 715-22, 1980.

[4] Welch R, Spatial resolution requirements for urban studies, IJRS vol.3, No.2, 1982.

[5] Kan E.P. et al, Data resolution VS forestry classification and modeling, Symposium on MPRS data, IB-24, 1975.

[6] Sadorusky F.A. et al, The influence of multispectral scanner resolutions on forest feature classification, Proc. 11th ISRSE, ERIM, 1977.

[7] Landgrebe P.A. et al, An empirical study ofscanner system parameters, IEEE Trans. GERS, 15, 1977.

[8] Kohei Arai, A consideration on an optimum spatial resolution of the multispectral scanner, 3rd Australasian RS Conference, 1984.

[9] Kohei Arai et al., Verification of Landsat MSS data based on land cover classification results, 8th Remote Sensing Symposium, 1982

[10] Townshend J. et al, Information extraction from RS data, A user view, IJRS, vol.2, No.4, 1981.

[11] Crapper P.F., The relation between pixel size land cover map accuracy, 3rd Australasian RS conference, 1984.

[12] Jackson M., Spatial analysis of TM products, Landsat-4 Early Results Symposium. 1983.

[13] Friedman H.D., On the expected error in the probability of misclassification, Proc. IEEE, vol.53, 1965.

[14] Kohei Arai, Classification by Re-Estimating Statistical Parameters Based on Auto-Regressive Model, Canadian Journal of Remote Sensing, Vol.16, No.3, pp.42-47, Jul.1990.

[15] Kohei Arai, Multi-Temporal Texture Analysis in TM Classification, Canadian Journal of Remote Sensing, Vol.17, No.3, pp.263-270, Jul.1991.

[16] Kohei Arai, Maximum Likelihood TM Classification Taking into account Pixel-to-Pixel Correlation, Journal of International GEOCARTO, Vol.7, pp.33-39, Jun.1992.

[17] Kohei Arai, A Supervised TM Classification with a Purification of Training Samples, International Journal of Remote Sensing, Vol.13, No.11, pp.2039-2049, Aug.1992.

[18] Kohei Arai, TM Classification Using Local Spectral Variability, Journal of International GEOCARTO, Vol.7, No.4, pp.1-9, Oct.1992.

[19] Kohei Arai, A Classification Method with Spatial Spectral Variability, International Journal of Remote Sensing, Vol.13, No.12, pp.699-709, Oct.1992.

[20] Kohei Arai, TM Classification Using Local Spectral Variability, International Journal of Remote Sensing, Vol.14, No.4, pp.699-709, 1993.

[21] Kohei Arai, Application of Inversion Theory for Image Analysis and Classification, Advances in Space Research, Vol.21, 3, 429-432, 1998.

[22] Kohei Arai and J.Wang, Polarimetric SAR image classification with maximum curvature of the trajectory in eigen space domain on the polarization signature, Advances in Space Research, 39, 1, 149-154, 2007.

[23] Hiroshi Okumura, Makoto Yamaura and Kohei Arai, A hybrid supervised classification method for multi-dimensional images using color and textural features, Journal of the Japanese Society of Image Electronics Engineering, 38, 6, 872-882, 2009.

[24] Kohei Arai, Polarimetric SAR image classification with high frequency component derived from wavelet multi resolution analysis: MRA, International Journal of Advanced Computer Science and Applications, 2, 9, 37-42, 2011.

[25] Kohei Arai Comparative study of polarimetric SAR classification methods including proposed method with maximum curvature of trajectory of backscattering cross section in ellipticity and orientation angle space, International Journal of Research and Reviews on Computer Science, 2, 4, 1005-1009, 2011.

[26] Kohei Arai, Comparative study on discrimination methods for identifying dangerous red tide species based on wavelet utilized classification methods, International Journal of Advanced Computer Science and Applications, 4, 1, 95-102, 2013.

[27] Kohei Arai, Multi spectral image classification method with selection of independent spectral features through correlation analysis, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 21-27, 2013.

[28] Kohei Arai, Image retrieval and classification method based on Euclidian distance between normalized features including wavelet descriptor, International Journal of Advanced Research in Artificial Intelligence, 2, 10, 19-25, 2013.

[29] Kohei Arai, Rosa Andrie Asmara, Gender classification method based on gait energy motion derived from silhouettes through wavelet analysis of human gait moving pictures, International Journal of Information Technology and Computer Science, 6, 3, 1-11, 2014.

[30] Kohei Arai, Rosa Andrie Asmara, Human gait skeleton model acquired with single side video camera and its application and implementation for gender classification, Journal of the Image Electronics and Engineering Society of Japan, Transaction of Image Electronics and Visual Computing, 1, 1, 78-87, 2013.

[31] Kohei Arai, Rosa Andrie Asmara, Human gait skeleton model acquired with single side video camera and its application and implementation for gender classification, Journal of the Image Electronics and Engineering Society of Japan, Transaction of Image Electronics and Visual Computing, 1, 1, 78-87, 2014.

[32] 503. Kohei Arai, Rosa Andrie Asmara, Gender classification method based on gait energy motion derived from silhouette through wavelet analysis of human gait moving pictures, International Journal of Information technology and Computer Science, 5, 5, 12-17, 2013.

[33] Kohei Arai, Rosa Andrie Asmara, Human gait gender classification using 3D discrete wavelet transformation feature extraction, International Journal of Advanced Research in Artificial Intelligence, 3, 2, 12-17, 2014.

[34] Kohei Arai, Image classification considering probability density function based on Simplified beta distribution, International Journal of Advanced Computer Science and Applications IJACSA, 11, 4, 481-486, 2020.

[35] Kohei Arai, Maximum Likelihood Classification based on Classified Result of Boundary Mixed Pixels for High Spatial Resolution of Satellite Images, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 9, 24-30, 2020.

[36] Kohei Arai, Context Classification based on Mixing Ratio Estimation by Means of Inversion Theory, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 44-50, 2020.

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

Disruptive Technologies for Labor Market Information System Implementation Enhancement in the UAE: A Conceptual Perspective

Ghada Goher[1], Maslin Masrom[2], Astuty Amrin[3], Noorlizawati Abd Rahim[4]
Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia

***Abstract*—In December 2019, the world learned about the first outbreak of the novel coronavirus (COVID-19) that first broke out in Wuhan, China. This limited outbreak in a small province of China has rapidly evolved into a global pandemic that has led to a health and economic crisis. As millions of individuals have lost their lives, others have lost their jobs due to the recession of 2020. While the skills and educational mismatch have been a prevalent problem in the UAE labor market, it is logical to assume that the global pandemic has likely increased this problem's extent. Therefore, there is an urgent need to adopt an agile, innovative solution to address the upcoming challenges in the labor markets due to the lack of skilled resources and the fear of future work amid the COVID-19 pandemic. Since industry and academia have identified skills and educational mismatch as a complex and multivariate problem, the paper builds a conceptual case from a system engineering perspective to solve this problem efficiently. Based on the literature reviewed related to disruptive technologies and labor market management systems, the paper proposes a new implementation approach for an integrated labor market information system enabled by the most widely used disruptive technologies components in the UAE (Machine Learning, AI, Blockchain, Internet of Things, Big Data Analytics, and Cloud Computing). The proposed approach is considered one of the immediate course of actions required to minimize the UAE economy's negative impact due to the presence of the skills and educational mismatch phenomena.**

***Keywords*—Disruptive technologies; labor market information systems; skills and educational mismatch; future of work; system engineering; system design thinking; COVID-19**

I. Introduction

The global COVID-19 pandemic has led to a complete transformation of the status quo and has challenged the current living and business landscapes. The ripple effects of the global pandemic hit all economic sectors in all countries. Under the current crisis, one aspect that has become clear is that there is a need to deploy disruptive technologies to solve common problems such as remote working, which exacerbated due to the need to keep the social distance. The current paper focuses on the skills and educational mismatch, one of the ever-lasting and continuously increasing problems in the UAE's labor market [1-3]. Since it is a large-scale problem, UAE policy-makers need to ensure that they are taking the appropriate steps towards disruptive technologies deployment while thinking about a solution to this problem. Despite that, the utilization of disruptive technologies penetrates the current processes and the paradigms that are followed [4]; the current practices of labor market information system (LMIS) implementation in the UAE has not yet leveraged such technologies. Therefore, proper utilization of disruptive technologies can potentially enhance the LMIS implementation practices and enable the UAE to solve the national mismatch problem for a better future of work. Such utilization will enable feasible development for an agile, innovative solution for such a complex problem and offer intelligent collaborations and workflows. From the other side, on a country level, labor market information systems are the most recommended solutions for better management of the labor markets as such solutions enable the identification and the collection of different types of information from the multi-stakeholders involved. With these in mind, the current paper is building a conceptual case to improve the future of work in the UAE labor market amid COVID-19 by potentially implementing an integrated labor market information system enabled by disruptive technologies.

The paper begins with a review of the concept of disruptive technologies and their various types while highlighting the opportunities that the technologies offer towards enabling the efficient development of complex solutions. Then, the paper presents an understanding of the Labor Market Information Systems (LMIS) and identifying their primary purposes for better labor market management. Following that, the paper portrays the mismatch outlook in the UAE labor market and the need for iLMIS implementation to resolve this issue. In addition, an explanation for the current trend for iLMIS implementation is provided, along with its limitations. Hence, a strategic course of actions is proposed to forward a better future of work in the UAE labor market amid COVID-19. Lastly, challenges and future guidelines will be discussed.

II. Disruptive Technologies

As all technology innovations can be considered disruptive, the list of disruptive technologies is increasing almost always. However, some technological innovations can transform the value pools and individuals' status quo in business and social landscapes. Therefore, countries need to be able to identify the scope of new technologies and the impact that they are going to have on people's lives and prepare accordingly. For this reason, it is rational to start the discussion with a quick review of the disruptive technologies. As identified by [4], disruptive technologies are classified as human-centric and smart-space-based technologies. Table I summarizes the stated types of disruptive technologies, along with their supporting emerging technologies, as outlined by [4].

TABLE I. CATEGORIZATION OF TECHNOLOGY

Disruptive technology theme	Disruptive technology types	Supporting emerging technologies
Human Centric	Hyper-automation	Machine Learning, Robotics
	Multi-experience	Virtual reality (VR), augmented reality (AR), and mixed reality (MR)
	Democratization	Big Data Analytics
	Human Augmentation	Virtual Reality, Augmented Reality, Virtual Assistants, Computer Vision, Biometrics
	Transparency and Traceability	Blockchain and Big Data Analytics, 5G
Smart Space	Empowered edge	Internet-of-things, Big Data Analytics, 5G
	Distributed Cloud	Cloud Computing, Big Data Analytics
	Autonomous things	Drones, autonomous vehicles, robotics, 5G
	Practical Blockchain	Blockchain
	AI Security	Artificial Intelligence, Blockchain

(Source: 4)

A. Human-Centric based Technologies

Human-centric technologies are one of the most transformative and impactful technologies at the individual level, whereby these technologies are intended to cater to the user's evolving needs. These technologies also define the interaction between users and the digital world. Some of the core types of these technologies include hyper-automation, multi-experience, democratization, human augmentation, as well as transparency and traceability [4].

Hyper-automation is a human-centric technology; it is a combination of several machine-learning techniques. It efficiently enables processes' automation and humans' augmentation. Moreover, it covers all the automation steps, including; discovering, analyzing, developing, implementing, monitoring, and re-assessing [5-6]. This type of technology includes two primary components: robotic process automation and intelligent business process management suites. The benefit of robotic process automation is enhancing the short-term processes and works by eliminating repetitive tasks using data manipulation and an integrated script structure that is tightly defined. Therefore, it facilitates the shift from legacy systems to new ones. While, the intelligent business process management suites help enhance the long-term processes and works as they also provide a unified solution that orchestrates between people, processes, and things. It is also a collection of tightly defined rules that can be used to oversee contextual work with its most common application to support the full life cycle of any processes end-to-end.

Furthermore, multi-experience technology provides users with different experiences and a new platform to interact with the digital world [4]. For example, technologies such as virtual reality, augmented reality, and mixed reality provide users with a multimodal sensory experience that can virtually deliver the relevant information and desired actions for different system users [7-8].

Besides, democratization is the technology that facilitates equally easy access to end-users (both potential and native) as well as expertise users (both discipline and domain) [4]. Such access facility is provided by using revolutionary technologies that can provide end-users with access to systems and information while needing minimal training. The technology of democratization uses artificial intelligence while developing a solution; this ensures that as time passes, the cost of the development of new solutions reduces and its production increases while supporting the developed solutions to run automatically based on the different users' interventions.

Human augmentation is defined as a technology that enhances human experience using physical and cognitive components. It offers humans many opportunities to develop their capabilities which contribute directly to a better condition. For instance, physical augmentation technologies provide humans with wearable devices that can enhance their physical bodies' capabilities, such as senses and biological functions [9]. On the other hand, cognitive augmentation technologies enhance human skills for better learning, development, and decision-making [10].

Lastly, transparency and traceability are defined as the ward that ensures trust, transparency, and ethics are addressed while disruptive technologies are used [11-12]. One of the core methods to restore trust is through paying attention to three areas while developing new systems powered by disruptive technologies: the use of AI, the use of personal data (privacy, control, ownership), and the use of a design that is ethically compliant [12].

Hence, Human-centric based Technologies are efficiently used in the case of complex systems development, especially in the case of customized profiling, distance collaboration between different users, and iterative enhancement of workflows are required.

B. Smart-Space based Technologies

Smart-space-based technologies connect different working environments and develop smart environments rich in information and provide context-specific intelligent services to different users [13]. It covers designing, implementing, and evaluating new information domains supported by new architectures to enable user-based intelligent services. There are several types of smart-space-based technologies such as edge computing, distributed cloud, autonomous things, practical blockchain, and AI security.

Edge computing is referred to as the technology that brings the computing and storage of data close to each other, significantly reducing bandwidth usage and enhancing response times when needed [14]. Therefore, it is predicted that approximately 50% of the companies will shift to edge computing from cloud computing [4].

The distributed cloud is also an advanced cloud computing technology that uses data storage at different geographic locations, enabling the distribution of service operations at these locations with the well-defined operation, governance, and evaluation of these distributed services. It operates under a tightly coupled system that has storage, computation, and networking capabilities. It is different from edge computing as

it is establishing a distributed cloud that provides closer data processing computing to the end-user, decreased latency, processed data, and enhanced security in a real-time environment.

On the other hand, autonomous things are inter-connected and AI-enabled physical devices. They are designed to replace humans; this includes robots, autonomous vehicles, and appliances. In essence, it is an enhanced level of automation as these devices deliver smart services to users. This technology's growth shall pave a pathway to shift from stand-alone intelligent systems to collaborative intelligent systems [15].

Moreover, a practical blockchain is a secure distributed ledger containing a chronological list of unalterable transactions records. These transactions can also be traced back to their origin, which makes them completely secure. Therefore, blockchain technology conveys trust, transparency, values and empowers collaborations across different digital eco-systems. However, blockchain is not scalable due to technical deficiencies and interoperability issues; thus, preventing it from being widely used in the business context [16].

Lastly, AI security. Based on the fact that artificial intelligence rapidly integrates into human decision-making and plays a crucial role in individuals' everyday lives, which has been facilitated due to the presence of the internet of things, cloud computing, and micro-services, along with the fact that individuals are increasingly connected. Therefore, with this level of penetration of artificial intelligence, there is a need to ensure that the safety, protection, and security defense of the AI-powered systems protect the individuals' data from cyber attackers.

Hence, smart-space-based technologies can be efficiently used as infrastructure in complex systems development, especially if required, for such development, to connect different working environments at different geographic locations to offer smart services with enhanced security to different users.

From amongst these advanced technologies outlined above, a total of six disruptive technologies have found widespread application in the UAE: Machine Learning, AI, Blockchain, Big Data, IoT, and Cloud Computing. The author in [17] noted that the UAE had established the groundwork for machine learning and AI learning by developing policies and providing incentives to drive their application. Furthermore, blockchain has also found widespread use in the UAE, with the public sector increasingly using blockchain to conduct its transactions [18]. Similar trends are noted in the UAE concerning big data analytics, whereby Dubai's Smart City Strategy is a big data-enabled program whereby big data analytics is poised to become a leading-edge technology in the UAE [19]. In addition, in terms of IoT, the application of this technology in the UAE is widespread, to extend that there is a new regulatory policy for the use of IoT in the country [20]. Finally, concerning cloud computing, an increase in usage is seen whereby there has been a surge in cloud computing use in the UAE in the past few years [21]. Therefore, these technologies that have widespread use in the UAE have been selected in this paper to build the conceptual case.

III. Labor Market Information Systems (LMIS)

A labor market information system is a set of foundations (government and private, workforce, employers, citizens) connected and collaborated to maximize the labor market's potential. Such effective management of the labor markets achieved as LMIS performs the identified roles and responsibilities for the production, storage, dissemination, and use of labor market data for better policy and program formulation and implementation.

Labor market information (LMI) is used as an umbrella term that represents all information related to the labor market, which includes the details of the labor supply and demand sides (structure, characteristics, and dynamics), as well as the information on lack of a labor market equilibrium. LMI exists in both "hard" (i.e., quantitative data and statistics) or "soft" (i.e., qualitative data on the functioning and characteristics of both labor market sides) forms.

In order to analyze the labor market and find the best information available regarding the state of the labor market, the collected raw data (both soft and hard) related to the labor market needs to be processed and transformed into what is called labor market intelligence [22, p. 4]. As illustrated in Fig. 1, LMI has three components; the input (labor market data), the process (labor market analysis), and the outputs (labor market intelligence).

Labor market information has four primary purposes:

1) Intervention-oriented, where LMI supports stakeholders to improve, resolve issues of the labor market.

2) Observation-oriented, where LMI serves labor market researchers to contribute to the economy and society's scholarly work.

3) Demand-oriented, where LMI increases employers' ability to hire the workforce efficiently.

4) Supply-oriented, where LMI offers a workforce, continues development to reduce the risk of exiting from the labor market.

Due to the complexity of the labor market management, the four purposes mentioned above of labor market information should be fulfilled; therefore, an LMIS must be comprehensive to provide value to all labor market stakeholders (i.e., policy-makers, researchers, employers, and workforce).

As observed by [23], this compressive LMIS are generally catering one of the two following functions as presented in Fig. 2.

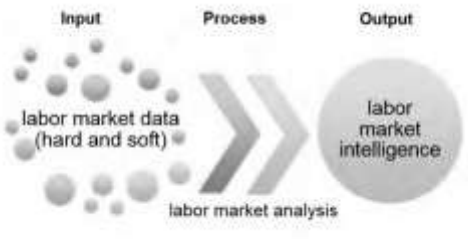


Fig. 1. The Components of Labor Market Information. (Source: 23).

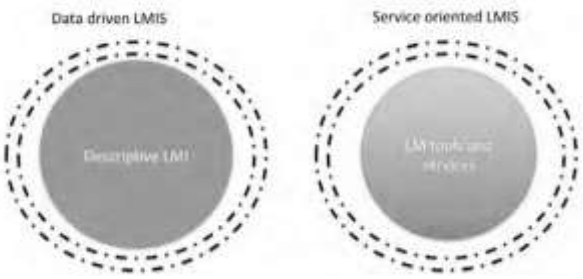


Fig. 2. Separated LMIS Functions. (Source: 23).

1) Data-driven LMIS: provides descriptive data on the labor market, mostly for the aim of intervention or observation. These systems build a set of statistical indicators such as unemployment rates, new job creation in a different sector, labor market demographics.

2) Service-oriented LMIS: provides labor market services to both sides of the labor market (demand and supply) to enhance their efforts to improve the work situation or the workforce, accordingly.

The author in [23] argued that instead of using separate functions of LMIS, an integrated LMIS (iLMIS), as illustrated in Fig. 3, should empower the labor market intelligence and analysis. Such integration for both functions shall identify more compressive LMIS implementation potentials representing the best solution for better labor market management, wherein data, policies, processes, and services are interlinked.

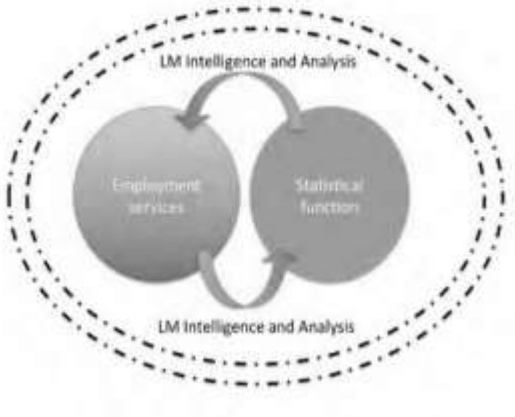


Fig. 3. Integrated LMIS (iLMIS). (Source: 23).

However, [23] did not consider the technology part for the feasible implementation of such a concept, iLMIS, neither the tools that allow for better gathering, processing, and disseminating labor market information. Hence, this paper proposes to complement this part by utilizing the wide applications offered by disruptive technologies.

A. Need for Integrated LMIS in the UAE Labor Market

Rapid technological advancements in the UAE have brought about a change in the status quo [24]. This technological change has not only replaced low-skill jobs but has also substantially transformed high-skill positions such as leadership, for instance. In other words, digital skills and knowledge of IT systems are crucial across all positions in an organization. In the UAE, which has transformed into the hub of rapid technological advancements and emerging technologies, skills and educational mismatch phenomena have become more prevalent in the labor market [25]. This phenomenon can negatively impact the country economy and employee-level outcomes [3], which is why there is a need to resolve the skills and educational mismatch problem.

Skills mismatch is defined as misalignment between the skills required by the employer and those possessed by the employee [26]. There are three types of skills mismatch: skill gap, shortage, and obsolescence. A survey revealed a predominant skills gap in the UAE labor market based on the research conducted by [3]. The same survey found a difference between what is expected from the demand side (the employers) and what is available on the supply side (the workforce). The author stipulated that one of the reasons for this difference was the lack of appropriate skills provision during the individual's education. Besides, this prevalence of skills gap in the UAE is deepened due to the rapid changes in the demand requirements; this is due to the new skills needed by the current unprecedented technological era, Industry 4.0, which is characterized by automation, advanced technology, and rapid change [24].

On the other hand, educational mismatch is defined as the incongruity between an individual's education and what is required by the job. The educational mismatch is generally of two types: horizontal and vertical mismatch. The horizontal mismatch is the difference between the individual's educational qualification and what is required by the job. While the vertical mismatch is the difference between the levels of educational qualification that an individual possesses and what is required by the job [27]. In the UAE context, [3] has identified a vast educational mismatch evident in the market, with horizontal mismatch being highly prevalent in the country relative to vertical mismatch. In other words, the authors have identified that there is a vast majority of individuals who are working in areas that are different from their field of education. One of the primary reasons that the authors identified to account for this educational mismatch in the UAE was that there is no alignment between what is required by the labor market demand (The employers) and what is supplied by the labor market suppliers (The higher education institutes).

Although the study by [3] offered an understanding of the predominance of skills and educational mismatch and its possible types in the UAE, there are limitations and disadvantages to the points highlighted as changes (seen and unseen) continuously reshape and unbalances the labor market. As a result, there is a need to understand the extent of skills and educational mismatch in the UAE and propose a feasible solution to solve this dynamic problem. Due to the COVID-19 pandemic, the overall frequency of job loss and economic downturn is much worse; therefore, the need to resolve the skills and educational mismatch becomes a matter of urgency to be fulfilled.

Therefore, to adequately address the highly prevalent skills and educational mismatch in the UAE's labor market and

effectively manage the unforeseen risks, there needs to develop a solution that can mitigate the vast and complex problem of skills and educational mismatch. This solution must consider a new relationship model between the known labor market's stakeholders while considering the working relationship's new norm, working from different geographic locations.

B. Current Trend for Implementing an Integrated LMIS

The implementation of iLMIS should benefit all stakeholders, individuals, businesses, educational institutions and government bodies. The primary purpose of the iLMIS implementation is to support the decision-making process for all stakeholders at different levels by ensuring that timely and relevant labor market information is provided. Hence, an implementation of iLMIS necessitates the integration, automation, and presentation of labor market information to all stakeholders. Therefore, the iLMIS implementation requires identifying infrastructure requirements, associated costs, governance needs, and potential features. In addition, the implementation needs to ensure that the iLMIS platform is deployed in a sustainable, secure, user-friendly, efficient, and easy to access way.

Fig. 4 illustrates a design-thinking framework that addresses the various general modules required for an iLMIS platform through the avenues of its integrated modules and data source owners.

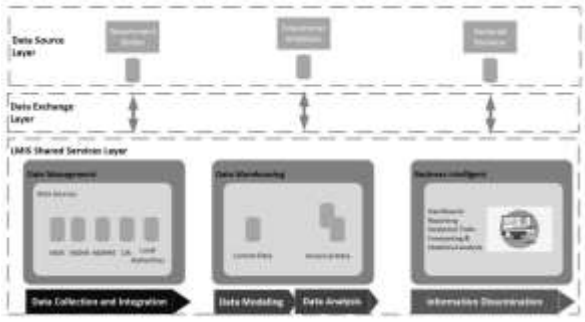


Fig. 4. Design Thinking Framework.

While Fig. 5 illustrates the ICT System Architecture of the current trend for an iLMIS infrastructure implementation.

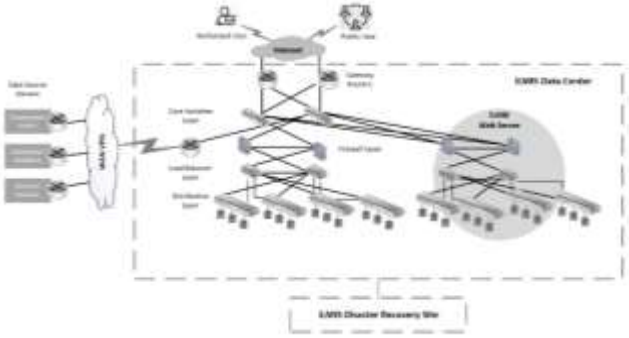


Fig. 5. ICT System Architecture of the Current Trend for iLMIS Infrastructure Implementation.

The current ICT infrastructure implementation approach focuses mainly on integration to route data from its sources through a secured wide area network (WAN-VPN) to be centrally hosted, analyzed, and interpreted. It requires the physical presence of a data center site (DC) as a focal point. DC is used to hosts applications, servers, security, load balancing, and data storage infrastructures. The data center is considered a hub that is accessed via the internet by all iLMIS users. A disaster recovery site (DR) is another physical location required for the recovery or continuity of the iLMIS operation in case of any disaster that happened to the DC, whether it is caused by natural or human. The iLMIS is implemented as a web-enabled platform that allows the authorized user and the job seekers to access the intranet systems at the DC, such as services, servers, applications, and databases.

One of the primary limitations of the current iLMIS implementation approach is that it does not consider the full lifecycle of the education-to-employment process end-to-end. There is a lack of orchestration between people, processes, and things. As shown in Fig. 4 and 5, there is increased fragmentation and lack of rightful ownership of the services, data, processes, and things. In addition to the lack of updated data from its sources, data owners are reluctant to share data. Therefore, it is challenging for such iLMIS implementation to understand proactive measures' principal characteristics without local and global data availability.

Furthermore, although the information or actions that different system users need do vary, the current iLMIS implementation approach does not provide a customized experience. Additionally, the current approach does not allow for developing a smart environment from different integrated environments. Moreover, in the current approach, it is not feasible to provide any real-time support to all stakeholders, as the computing and data storage do not co-exist in the same location. Also, the current approach does not support the transaction records' traceability, which limits the collaboration between stakeholders. Finally, the current approach does not facilitate smart, fast, and analytics-driven decision-making data.

Therefore, it is required to discover an efficient approach that supports a complex infrastructure implementation of an iLMIS. The new approach needs to enable connecting different working environments, distance collaboration, interactive workflow, intelligent services, and customized profiling.

C. The Solution to a Better Future of Work

Since the COVID-19 pandemic has started, disruptive technologies such as artificial intelligence, cloud computing, and the internet of things have been used for many purposes [28]. The author has outlined the use of disruptive technologies for saving human lives, increasing the understanding of the novel coronavirus disease, and enabling humanity to defeat the virus [28]. Furthermore, artificial intelligence has been called in for crisis management and national emergency during the pandemic. Besides, disruptive technologies have also been used to tackle sources of misinformation and generate reliable scientific information, as well as for assisting researchers in genome sequencing of the virus [29-31].

This widespread use of disruptive technologies can be extended to mitigate the skills and educational mismatch problem. The phenomenon of skills and educational mismatch is highly complex and multifaceted and hence, requires an equally powerful and complex solution structure that targets all the factors that are increasing skills and educational mismatch in the labor markets [3]. The study by [3] served as an exploratory study and did not provide a comprehensive insight into the possible solution for mitigating skills and educational mismatch problems. However, the authors state that emerging technologies need to be considered in developing a solution, which can dynamically reduce the mismatch between the demand and supply sides of the UAE labor market.

The author in [23] identifies solutions that address the study's limitation by [3]. In addition, [23] made a case for the iLMIS's values which are achievable if employment services and statistical functions are allowed to interact and dynamically benefit from each other. However, [23] did not look at the proposed solution's architecture, iLMIS, from its technological perspective for feasible implementation. The study was limited to a statement highlighting the need to identify the technological components that ensure offering user-centric labor market information and services to all stakeholders. This limitation is addressed in the current paper by proposing the utilization of disruptive technologies components in the implementation of iLMIS. The offered components cover the technology part of implementing the iLMIS; this includes IT infrastructure, data collection tools, and technologies for processing and disseminating labor market information.

It is necessary to mitigate the skills and educational mismatch problem in the UAE as it can lead to employee productivity issues, wage dissatisfaction, and other unfavorable labor market outcomes [32-34]. Besides, another fact that the COVID-19 pandemic has further deepened this problem in UAE and other countries across the globe. It is required to promptly provide a new approach to manage the current labor market eco-system effectively and elevate its role as the main component of its economic development planning over the coming few years. The new approach should consider connecting all the labor market's stakeholders, the workforce and representatives of the demand, the supply, and the regulator sides, as the data exchange mechanisms would include input from them. The collected data shall be processed and analyzed with only the relevant information being extracted and targeted to the end-users.

Fig. 6 illustrates a system design architecture that complements the [23] study's outcomes and proposes implementing an iLMIS smart platform enabled by advanced technological components outlined in the earlier discussion (machine learning, AI, blockchain, big data analytics, and cloud computing).

By adopting the proposed implementation approach, all the labor market stakeholders will be connected dynamically through a complex, intelligent, and automated system, as depicted in Fig. 6. The proposed iLMIS smart platform oversees the entire life cycle of the education-to-employment journey, ensuring more splendid orchestration between various stakeholders enabled by machine learning. It will also provide a new customized experience, information, and actions needed for different users as enabled by big data analytics. In addition, the proposed iLMIS smart platform will operate as a smart environment that is developed through the integrating of various environments using the power of IoT. Moreover, the smart platform will also support real-time stakeholder response, data computing, and data storage using cloud-computing technologies. Furthermore, as blockchain technologies will enable the smart platform, the transaction records' traceability will be facilitated, extending the collaboration between stakeholders efficiently. Finally, as AI will enable the smart platform, smart decision-making will be facilitated, and decisions driven by real-time data will be carried out.

Therefore, in line with the insights presented above, the current paper proposes a strategic course of actions to solve the UAE labor market's mismatch problem. The UAE may consider a better way to manage the labor market by implementing an iLMIS supported by disruptive technologies (Machine Learning, AI, Blockchain, Internet of Things, Big Data Analytics, and Cloud Computing). Hence, the implemented iLMIS solution can effectively handle the upcoming challenges in the UAE labor market amid the COVID-19 pandemic. This course of actions shall deliver at a low cost a feasible solution that is easy to implement, efficient, reliable, robust, and user-friendly. In addition to that, it shall allow all stakeholders to interact and dynamically benefit from each other. As a result, the iLMIS shall offer more adapted services to suit better employers and workers, as well as the data collected through the management of services to construct more accurate and predictive models of the UAE labor market performance. Table II outlines the SWOT analysis for the proposed iLMIS implementation approach towards reducing the skills and educational mismatch evident in the UAE labor market.

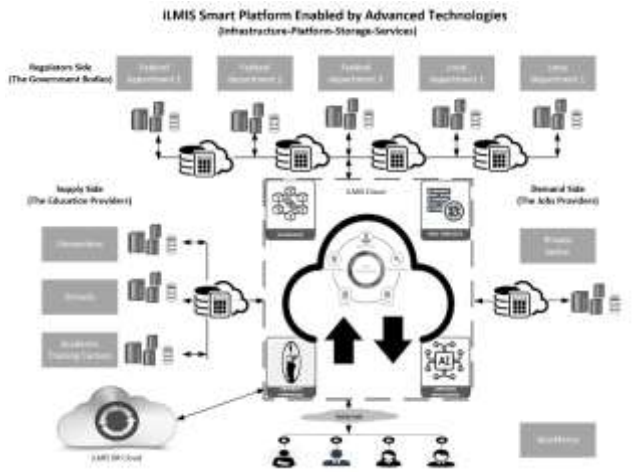


Fig. 6. System Design Architecture.

TABLE II. SWOT ANALYSIS FOR THE PROPOSED STRATEGIC COURSE OF ACTIONS TOWARDS REDUCING THE SKILLS AND EDUCATIONAL MISMATCH EVIDENT IN THE UAE LABOR MARKET

SWOT	Remarks
Strength	1. Mismatch phenomena discovery and management in a cost-effective and time-efficient approach. 2. Quick prediction and propagation analysis of the mismatch phenomena while limiting human intervention. 3. On-time detection of the demand-supply imbalance as it happens to result in faster analysis and prompt corrective actions. 4. It is a fast, cost-effective, and self-diagnosis approach for the root causes of the dynamic mismatch phenomena, enabling stakeholders to focus their operation and resources while contributing to the solutions. 5. It leads to discovering permanent, cost-effective solutions for the mismatch phenomena and contamination of its spread. 6. Supports lessens the economic buzzes regarding labor market issues with automated approaches using artificial intelligence. 7. Offers time-efficient and automated responses that are available 24x7 for all stakeholders, which endure; secure, dynamic, and robust mismatch solution and prevention mechanisms. 8. Smart problem-solving techniques ensure efficient utilization of the interdependent systems' resources and effective diagnosis without affecting individual stakeholders' operations. 9. Provides effective monitoring of the current status of the mismatch phenomena through; ensuring gathering the latest updates and integrating with several heterogeneous sources to a single platform. 10. Effective data exchange mechanism ensures the availability of all essential data to all stakeholders as needed.
Weakness	1. Relying heavily on the current UAE labor market information patterns may generate biases concerned with the incoming drifting data. 2. Proposed disruptive technologies approaches require heavy computation resources to detect the mismatch phenomena that might not be available to the UAE labor market stakeholders. 3. Wrong pattern discovery may lead to severe results in the real national labor market; therefore, the high quality of data is required to avoid the spread of false information and the generation of unnecessary economic fear situations. 4. The required resources to design, deploy and operate the solution may not be available in the country. 5. The solution is challenged, as the processing of the official language, Arabic, is required to recognize fake news. 6. Several trials would consume a massive amount of resources, efforts, time, and financial support for such an innovative way. 7. Solution adaptability is a critical issue for national governing bodies as it requires a central-management, overarching infrastructural which changing the current roles and responsibilities of the individual shareholders. 8. The involvement of all labor market stakeholders and their technological awareness is needed for more accurate solution results. 9. The solution depends on the authorized sources of information; hence chances of duplicate information are high and micro-level analysis is required for all the labor market stakeholders' data. 10. The solution may raise security concerns and be incompatible with the current labor market stakeholders' IT infrastructures and systems.
Opportunities	1. The prediction and propagation analysis will minimize the amount of economic loss based upon the current situation of the mismatch phenomena in the UAE and lead to the generation of the follow-up plans to mitigate the early planning issues containing its spread. 2. Disruptive techniques support the central analysis procedure with automated approaches that promptly detect the mismatch phenomena and spare more time to assess alternative solutions instead of distributed and interlinked analysis. 3. AI-based approaches for solution discovery can be executed without causing any harm to the country's economy. 4. Provides a chance to the labor market stakeholders to focus more on strategies, policies, and legislations making for proper operation of the complex solution and encourages them to develop better behaviors concerning cross-work and data sharing to achieve common goals and targets. 5. Protects from economic infodemics by implementing better learning algorithms for more robust innovative solutions and monitoring mechanisms for the mismatch phenomena. 6. Fortifies the world economies with more promising and robust solutions for the ever known mismatch phenomena and encourages 0% unemployment rate economy and opening new avenues for; smart-space infrastructure development, labor relation standard and rights, smart working environment, automated surveillance with limited human resources. 7. It increases the effectiveness and transparency through; technological advancements, platform compatibility, and micro-level analysis. 8. It permits the high potential for broader application across various countries. 9. Will allow companies to gain cross-country access to hiring individuals remotely with relevant and matched skills and vice versa. 10. The disruptive technologies have proven record for being feasible and thriving in implementing and operating such solutions.
Threats	1. Security concerns related to sharing private information globally with the possibility of false-negative predictions could generate an unnecessary panic situation. 2. The AI approaches have a percentage of error; therefore, false detection of the mismatch phenomena could fire back, and hence human supervision of the results is required before sharing labor market information in the public platform. 3. The solution's discovery simulations may deviate from the real world and cause a national economic disturbance. 4. The shared information may raise the anxiety level among individuals or organizations impacted by unwanted executed actions, and as a result, they may share fabricated information on public platforms that could mislead the nation. 5. The high acceptance resistance to handling the mismatch phenomena via technology-based solutions even if the solution shall provide information of public concern. 6. Novel discovered actions, untested, might generate undesirable impacts on the economy. 7. The technology-based readiness of economic disruption may increase the unemployment ratio. 8. There is a potential for too much dependency on the technologies, and hence malfunctioning of such solution, privacy, trust, and security issues become significant concerns. 9. Demand prediction of the required workforce skills is the most crucial issue, as due to false prediction, the mismatch phenomena will continue coming into the picture. 10. False indicators may cause unnecessary economic disruption; therefore, the solution, design, implementation, supervision, and maintenance should be handled with profound attention.

IV. Challenges

Several challenges can limit implementing of the proposed strategic course of actions, iLMIS enabled by disruptive technologies, to solve the skills and educational and educational mismatch problem in the UAE.

One of the core challenges is related to regulatory challenges whereby the use of disruptive technologies in the labor market sector to resolve skills and educational mismatch should be considered carefully with the UAE's regulatory laws. While these technologies bring benefits, it also poses legal and regulatory issues as, currently, the accountability of the education-to-employment eco-system is distributed across multiple parties. For example, in the case of using blockchain technologies, it will be essential to consider what law might apply and the potential risks to be mitigated. Similarly, in the case of using AI, creating national standard norms of legal and governance models may be considered more contented to dictate AI usage in the UAE labor market. In addition to that, government regulators' privacy considerations, private organizations, and individuals might prevent access to data, which is needed for the proposed strategic course of actions to be successful.

Another challenge to implementing iLMIS technology-enabled is that there is no unified data center in the UAE or a linked structure that can provide access to information required from or for the parties involved. A lack of central unified databases is a critical challenge to solve the issue of mismatch. Most of the current labor market information databases are collected from several sources; however, it is not sufficient for iLMIS to create more significant impacts to solve a complex problem. On a broader scale, most of the time, countries are reluctant to share data; therefore, it is challenging to understand proactive measures' principal characteristics without local and global data availability.

V. Future Guidelines

The future guidelines are based on the unanticipated risks that affect the world economies in 2020 due to the COVID-19 pandemic, which seems to be an unending pandemic. Thus, there is an urgency to accelerate liable action plans to overcome the deeper mismatch problem during the crisis, have a solution for this problem, and ensure labor equilibrium for a better UAE economy for the generations to come. This complex problem demands a severe contribution from policy-makers and regulators to demonstrate legal plausibility, ethical soundness, and effectiveness of deploying a national iLMIS enabled by emerging, future, and disruptive technologies under the current time pressure. The accelerated action plan needs to consider three main stages that must be executed consecutively: acknowledging, reinforcing, and re-engineering.

The first stage is acknowledging. It is essential to admit the prevalence of the mismatch problem in the national labor market and act decisively to have the best possible solution. Such a complex solution is feasible to be implemented and operated based on the existing disruptive technologies to resolve the skills and educational mismatch issue amid the COVID-19 pandemic. This stage shall support continuity, productivity, and prosperity of the UAE's economy and gain the buy-in of the labor market's stakeholders to move towards the success of such a solution.

The second stage is reinforcing. Once acknowledgment of the problem and the need for an urgent solution is realized, the reinforcing stage needs to be covered. It is required to have sufficient time to identify the structure of the iLMIS solution and the disruptive technologies components required for its implementation and operation. It is equally vital to realign priorities and allocate resources to assess each of the emerging technologies to ensure a balance between costs and benefits while identifying the technologies to be used to build the new solution (iLMIS). The next step is to strengthen and build the identified technological capabilities that will lead to efficient automation for the new solution (iLMIS) processes and workflows that encounter the mismatch problem challenges amid COVID-19. It is indicated that stand-alone technologies cannot add value by themselves; however, they can play a prominent role in implementing and operating a solution. Therefore, it is crucial to identify the possible new solution (iLMIS) topology and the needed disruptive technological solutions in the labor market context. Moreover, it is equally essential to ensure proper design, policies, and regulatory measures are put in place for a successful implementation of this solution.

The third stage is the re-engineering. Once a prompt response to solve the problem is taken, the re-engineering stage is initiated. The new solution (iLMIS) shall impose genuine and dynamic collaboration between the labor market's stakeholders to change their current operations models. It is mandatory to develop a framework that leads the shift from the current eco-system to the new eco-system powered by the new solution (iLMIS). Serious effort is required to review and assess the current eco-system and its technology-based solutions to develop the new solution (iLMIS). This effort must include the current strategies, plans, infrastructure, systems, laws, rules, and guidelines to reframe the existing setups and design the new setups to resolve the stated issue. After achieving that, proper time needs to be allocated to drive rapid growth and sustainable development of the new solution (iLMIS) to gain momentum in all labor market sectors and steer the economic wheel towards a better future of work in the UAE labor market. This stage shall support guiding the successful implementation of the new solution (iLMIS).

VI. Conclusion

The paper first introduced the two categories of the concept of disruptive technologies (human-centric and smart-space). This introduction was followed by an explanation of the LMIS concept and its types. Then the paper has portrayed the mismatch outlook in the UAE labor market amid COVID-19 and the need for iLMIS to resolve this issue. The followed section highlighted the current trend for iLMIS implementation and its limitations. Hence, a proposition was made for implementing an iLMIS enabled by disruptive technologies components as one of the immediate courses of action to solve the mismatch problem in the UAE labor market. The same section included a SWAT analysis for the proposed approach. The paper then covered the foreseen challenges found at two levels in the UAE labor market; regularity and decentralized

ownership of data. Lastly, future directions have been presented to highlight research areas within this body of knowledge.

The implementation approach proposed is a novel contribution to resolve the issue of skills and educational mismatch evident in the UAE labor market and grew deeper amid COVID-19. The paper highlights the potential for leveraging six disruptive technologies in implementing an iLMIS to provide a dynamic linkage and facilitate orchestration between and across various labor market stakeholders' groups. The leveraging of disruptive technologies towards solving the mismatch problem was mostly missing from past researches, and the current study bridges this gap. The proposed approach is dictated by the complex nature of the skills and educational mismatch problem and the urgent need for efficient development of 21st-century skills. It is believed that with the vast amount of data required to resolve the aforementioned problem, which is sourced by multi-stakeholders, the need for using iLMIS enabled by disruptive technologies is necessary to facilitate its mitigation.

REFERENCES

[1] A. Tamimi and C.-E. Kressner, "The Gap between Education, Talent, and Technology in the UAE | Lexology," www.lexology.com, 2016. https://www.lexology.com/library/detail.aspx?g=05fd91cd-45d8-4026-9fde-ec7ab5fa2075 (accessed 22 February 2021).

[2] British Council, "Future skills supporting the UAE's future workforce," 2014. [Online]. Available: https://www.britishcouncil.ae/sites/default/files/bc_futureskills_english_1mar18_3.pdf.

[3] Goher, G., Masrom, M., Amrin, A., & Abd Rahim, N. "UAE's Labor Market Snapshot | Skills and Educational Mismatch During Industry 4.0", International Journal of Recent Technology and Engineering, vol. 8, pp. 1332–1340, 2020. https://doi.org/10.35940/ijrte.f7626.038620.

[4] Gartner, 2020. "Top 10 Strategic Technology Trends for 2020". 2020. [Online]. Available: https://www.gartner.com/en/publications/top-tech-trends-2020.

[5] van der Aalst, Martin Bichler, and Armin Heinzl, "Robotic process automation", 2018.

[6] Dunie, R., Schulte, W. R., Cantara, M., & Kerremans, M, "Magic Quadrant for intelligent business process management suites", 2020.

[7] Farshid Mana, Jeannette Paschen, Theresa Eriksson, and Jan Kietzmann. 2018. Go boldly!: Explore augmented reality (AR), virtual reality (VR), and mixed reality (MR) for business. Business Horizons 61, 5 (2018), 657–663.

[8] Milgram Paul and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. IEICE TRANSACTIONS on Information and Systems 77, 12 (1994), 1321–1329.

[9] Crowell Philip, Gregory Kanagaki, Meghan O'donovan, Courtney Haynes, Joon-hyuk Park, Jennifer Neugebauer, Edward Hennessy, Angela Boynton, Blake Mitchell, Andrew Tweedell, and Henry Girolamo. "Methodologies for Evaluating the Effects of Physical Augmentation Technologies on Soldier Performance Methodologies for Evaluating the Effects of Physical Augmentation Technologies on Soldier Performance," 2018.

[10] Cinel Caterina, Davide Valeriani, and Riccardo Poli. "Neurotechnologies for Human Cognitive Augmentation: Current State of the Art and Future Prospects." Frontiers in Human Neuroscience, vol. 13, 2019.

[11] Veloso Bruno M, Fátima Leal, Benedita Malheiro, and Juan Carlos Burguillo, "A 2020 perspective on Online guest profiling and hotel recommendation: Reliability, Scalability, Traceability and Transparency". Electronic Commerce Research and Applications vol.40, 2020.

[12] Haroon, M. Basharat, A. M. Khattak, and W. Ejaz, "Internet of Things Platform for Transparency and Trace-ability of Food Supply Chain." In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 0013–0019, 2019.

[13] El-Fakih Khaled, Teruhiro Mizumoto, Keiichi Yasumoto, and Teruo Higashino, "Energy aware simulation and testing of smart-spaces." Information and Software Technology, vol.118, 2020.

[14] Hongjing Ji, Osama Alfarraj, and Amr Tolba. 2020. Artificial Intelligence-Empowered Edge of Vehicles: Architecture, Enabling Technologies, and Applications. IEEE Access 8 (2020), 61020–61034.

[15] Langley David J, Jenny van Doorn, Irene CL Ng, Stefan Stieglitz, Alexander Lazovik, and Albert Boonstra. 2020. The Internet of Everything: Smart things and their impact on business models. Journal of Business Research (2020).

[16] Fabiano Nicola. 2017. Internet of Things and blockchain: legal issues and privacy. The challenge for a privacy standard. In 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE. 727–734.

[17] S. Araz, "The UAE eyes AI supremacy: A key strategy for the 21st century," Middle East Institute, 2020. https://www.mei.edu/publications/uae-eyes-ai-supremacy-key-strategy-21st-century (accessed 22 February 2021).

[18] UAE Government, "Blockchain in the UAE government - The Official Portal of the UAE Government," u.ae, 2021. https://u.ae/en/about-the-uae/digital-uae/blockchain-in-the-uae-government#:~:text=The%20UAE%20Government%20adopted%20blockchain%20technology%20in%20conducting%20its%20transactions.&text=The%20Emirates%20Blockchain%20Strategy%202021%20aims%20to%20capitalise%20on%20the (accessed 22 February 2021).

[19] Deloitte, "Big Data in the GCC," Deloitte, 2019. https://www2.deloitte.com/ye/en/pages/about-deloitte/articles/revolution/big-data-gcc.html (accessed 22 February 2021).

[20] B. McKenzie, "UAE Regulates Internet of Things | Insight | Baker McKenzie," www.bakermckenzie.com, 2019. https://www.bakermckenzie.com/en/insight/publications/2019/05/uae-regulates-internet-of-things (accessed 22 February 2021).

[21] A. Buller, "Cloud computing surges in the UAE in 2019," ComputerWeekly.com, 2019. https://www.computerweekly.com/news/252475177/Cloud-computing-surges-in-the-UAE-in-2019 (accessed 24 November 2020).

[22] Green, F. and Zhu, Y, "Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education," Oxford economic papers, vol.62, pp.740-763, 2010.

[23] National Skill Development Corporation (NSDC). "Concept paper on labor market information system: an Indian perspective," 2011. http://www.tsscindia.com/media/1433/concept-paper-on-lmis.pdf.

[24] Sparreboom, T., & Powell, M. Labor market information and analysis for skills development (No. 994340413402676). International Labor Organization. 2009.

[25] J. Connell et al., "The importance of content and face validity in instrument development: lessons learned from service users when developing the Recovering Quality of Life measure (ReQoL)," Quality of Life Research, vol. 27, no. 7, pp. 1893–1902, Apr. 2018, DOI: 10.1007/s11136-018-1847-y.

[26] Khaleej Times, "49% of UAE respondents believe 'skills gap' exists," Khaleej Times, 2017. https://www.khaleejtimes.com/49-of-uae-respondents-believe-skills-gap-exists (accessed 14 February 2020).

[27] McGuinness, S., Pouliakas, K., & Redmond, P. "How useful is the concept of skills mismatch?." 2017.

[28] Di Stasio, V. "Diversion or safety net?' Institutions and public opinion on vocational education and training". Journal of European Social Policy, vol. 27, pp.360-372, 2017.

[29] Nguyen, T. T., "Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions". Preprint, vol.10, 2020.

[30] Jumper, J., Tunyasuvunakool, K., Kohli, P., Hassabis, D. and the AlphaFold Team (2020). Computational predictions of protein structures associated with COVID-19, DeepMind website, 5 March 2020,

https://deepmind.com/research/opensource/computational-predictions-of-protein-structures-associated-with-COVID-19.

[31] Etzioni, O., & DeCario, N. "AI can help scientists find a COVID-19 vaccine". WIRED. 28 March 2020.

[32] Grunau, P., "The impact of overeducated and undereducated workers on firm-level productivity–First evidence for Germany. In Forthcoming, Institute for Employment Research (IAB), Nuremberg, Germany". Paper presented at the Workshop Firm-level Analysis of Labor Market Issues, Université Catholique de Louvain, Belgium, 2014.

[33] Mahy, B., Rycx, F. and Vermeylen, G., "Educational mismatch and firm productivity: Do skills, technology and uncertainty matter?" De Economist, vol.163, pp.233-262, 2015.

[34] Allen, J. and Van der Velden, R., "Educational mismatches versus skill mismatches: effects on wages, job satisfaction, and on-the-job search," Oxford economic papers, vol. 53, pp.434-452, 2001.

AUTHORS' PROFILE

Eng. Ghada Goher is a researcher Ph.D. student at Razak Faculty of Technology and Informatics, previously known as Razak Faculty of Engineering and Advanced Technology, Universiti Teknologi Malaysia. Because she believes that "Life itself is your teacher, and you are in a state of constant learning. Bruce Lee", she is starting her researching work with 27+ years of comprehensive large-scale management experience in diversified areas of the ICT Industry. She is known for developing strategies, policies, procedures, and operation models. She has proven records of accomplishment of government services providing model excellence. She was a pioneer in integrating the franchising techniques with the government service providing models. She possesses substantial experience in UAE and Egypt, specifically in the following business areas; Immigration, Border Control, Labor, Health, Airline, Banking, and Courts. She holds an Engineering Bachelor, Software Development diploma, Technical Professional Diploma, MCSE Certification, PMP Certification, Information Technology MSc, and Master of Business Administration (MBA). She was a recipient of various awards such as; The Technical Excellency of the year from Microsoft Egypt "System Architect." She has been an active member of the Project Management Institute (PMI), Arabian Gulf Chapter, and UAE Chapter.

Prof. Dr. Maslin Masrom is an Associate Professor at Razak Faculty of Technology and Informatics, previously known as Razak Faculty of Engineering and Advanced Technology, Universiti Teknologi Malaysia (UTM). She received her Bachelor in Computer Science from Universiti Teknologi Malaysia. She received her Master of Science in Operations Research from Western Michigan University, the USA, and a Ph.D. in Information Technology/Information System Management from Universiti Putra, Malaysia. Her teaching experiences have been focusing on operations research/operations management, IT/IS management, knowledge management, and ethics in computing. Her current research interest includes it-adoption, e-government, technology management, information security management, women and technologies, cloud computing, structural equation modeling, and creativity and innovation management. She was a recipient of various awards, including Best Paper Award, Knowledge Management International Conference, Best Publishing Award Multimedia and Communication Technology, McGraw-Hill Publishing Best Paper Award in Operations Management / Management Information Systems, Universiti Teknologi Malaysia Excellence Service Award, Universiti Teknologi Malaysia Distinction Service Award. She was appointed as a Visiting Professor at many International Engineering and Technology universities. She has chaired numerous technical programs and has been an active member of various national and international technical committees.

Prof. Dr. Astuty Amrin is an Associate Professor at the Department of Engineering, Razak Faculty of Technology and Informatics, previously known as Razak Faculty of Engineering and Advanced Technology, Universiti Teknologi Malaysia (UTM). Currently, she is also the Dean of Razak Faculty of Technology and Informatics UTM. She received her Bachelor in Materials Eng. (pioneer batch) from Universiti Sains Malaysia, MSc in Corrosion Sc. and Eng. from UMIST, UK and Ph.D. in 2005. Her teaching experiences have been focusing on technology management, creativity, and innovation management, maintenance management, research methodology, innovation & new product development, materials science, and technology. She was a recipient of various awards, including Award of Excellence for Active Blended Learning in Technology Management course. Her research interest is devoted to Materials Engineering, specifically on high-temperature oxidation of alloys, compositional modification of newly developed (α+β)Ti-Alloys, developing accelerated corrosion test procedures for Malaysia automotive industry, ultraviolet treatment for microbially-influenced corrosion of steels, rejuvenation of Ni-Cr superalloy turbine blade. She was appointed as a Visiting Professor at Sudan University of Science and Technology in February and December 2016. She has also been appointed as a Visiting Professor at the King Mongkut's University of Technology Thonburi, Bangkok, Thailand, since 2015 and panel of experts for MSc in Corrosion Engineering program for Universiti Teknologi Petronas. She has chaired numerous technical programs and has also been an active member of various national and international technical committees, advisory boards, program committees, and editorial boards.

Dr. Noorlizawati Abd Rahim is a senior lecturer at the Science, Management & Design Department, Razak Faculty of Technology and Informatics, previously known as Razak Faculty of Engineering and Advanced Technology, Universiti Teknologi Malaysia. Her teaching focuses on entrepreneurship, quantitative data analysis, and semiconductor materials engineering. Her research interests are in the areas of technology, entrepreneurship, and entrepreneurship education. She is a member of the IEEE Technology and Engineering Management Society and Malaysia Nanotechnology Association. Before the faculty appointment, she had industrial experiences in chipset design and development, semiconductor manufacturing, and technology commercialization from Intel, Freescale Semiconductor, and NanoMalaysia. She received her BEng in Electrical & Electronic Engineering from Cardiff University, MSc in Nanotechnology from University College London (UCL), and Ph.D. in entrepreneurship from Universiti Teknologi Malaysia. Noorlizawati was the recipient of the 2017 Innovation Book Award by Malaysian Technology Development Corporation, the 2012 MSc Nanotechnology Achievement Award from UCL, and the 2008 IET Electrical & Electronic Engineering Prize.

Exploratory Study of Some Machine Learning Techniques to Classify the Patient Treatment

Mujiono Sadikin[1], Ida Nurhaida[2]
Faculty of Computer Science
Universitas Mercu Buana, Jakarta, Indonesia

Ria Puspita Sari[3]
Siloam Hospital
Jakarta, Indonesia

***Abstract*—Numerous studies have been carried out on computation and its applications to medical data with proven benefits for improving the quality of public health. However, not all research results or practical applications can be applied to all conditions but must be in accordance with the various contexts such as community culture, geographical, or citizen behaviors. Unfortunately, the use of digital data in Indonesia is still very limited. The study objective is to assess various data mining techniques to utilize data from laboratory test results collected from a private hospital in Indonesia in predicting the next patient treatment. Furthermore, various machine learning classification techniques were explored for the purpose. Based on the experiments, it was concluded that XGBoost with hyperparameter tuning produced the best accuracy level at 0.7579, compared to other classifiers. A better level of accuracy can be obtained by enriching the type of dataset used, such as the patient's medical record history.**

***Keywords*—Electronic health record; XGBoost; patient treatment; patient laboratory test data**

I. INTRODUCTION

The success of medical treatment services is dependent on the quality of health services and the information precision related to the medical aspects [1]. Unfortunately, the access performed to the relevant medical information is increasingly difficult due to the rapid growth in data volume and its heterogeneous format as well. Health care is one of the most complex industries which includes many stakeholders, various tools, and technologies as well [2]. The new techniques are always needed to assist the dealing with this type of data with the computational technique used to address the problem related to medical information.

Various techniques have been carried out on medically related datasets for many purposes. Due to the broad scope of the medical and health fields, various research topics are found which ranges from diagnosis [3]–[5], diseases [3], [4], [6]–[11], patient's condition [12]–[17], prescription and medication [1], [18]–[21], to genomics [22]–[24].

The development of the health system, its problems, and challenges tightly relates to multi factors and contexts such as geographic location, local regulation, community-style demographics, wealth level, etc. The contextual factors are the most important part used to develop the health-medical researches endorsed by the Agency for Healthcare Research and Quality USA [25]. These factors are following the World Health Organization, which supports the achievement level of best health services quality, and medical devices operation base on the contextual context [26]. The studies regarding the contextual factors in developing the Primary Health Center (PHC) [39], showed that many factors such as social models, an institutional context that promotes risk-averseness and patient care, infrastructure, community expectation, and doctors' disinterest in primary care roles need to be considered.

Unfortunately for our local context i.e. Indonesia, compared to a very large population and a very large area of the country, the studies regarding the medical records or electronical health are very limited. Some of the studies which focus on the local context are published in [11] and [15]. The first article presents the study results of early dengue disease detection with the dataset captured from some public health (PUSKESMAS). The study overcame the problem associated with physical detection methods in detecting the patient's symptoms by comparing some conventional classifiers with the ELM technique. In the second study, the authors performed the toddler's nutritional status identification using the clustering method, which is categorized into 5 clusters: good, moderate, malnutrition, over, and obesity. The other study is the enrichment of ontology in tuberculosis epidemiology domain use the pulmonary TB (Tuberculosis) scientific documents [27].

Considering the importance of the right context in conducting health-medical research, this study was conducted to utilize the value of patient medical data from the results of laboratory tests taken from one of a private hospital in Indonesia. This dataset is the main consideration factor used by doctors to determine the next course of action for patients, whether they need to be hospitalized (in-patient care) or not (out-patient care). Various literary studies show that AI and Datascience-based tools are proven to be able to improve the quality of health services. According to authors' knowledge, in the field of health care in Indonesia, there are no AI-based tools available. This research is an attempt to contribute to this field.

In this study, we elaborate on some of the machine learning techniques used to classify these patient treatments based on the laboratory test results data. Compared to the other technique, the XGBoost with Grid Search hyperparameters optimization performance is outperform.

In addition, this research utilized EHR data from the local context to determine the characteristics of patients' treatment, with similar pattern distribution. Therefore, the machine learning technique was proposed by the authors to handle this problem. The article is organized as follows: the first section

describes the background and justification of the research. This is followed by section two; the material and method of study are used to analyze the dataset and research methodology. The next section presents the results and discussion regarding the experiments. And in the last section, we summarize our research findings as a study conclusion.

II. Related Work

Artificial intelligence (AI) – Machine Learning (ML) is one of the most powerful techniques used to address any problem in the health and medical sector. The study carried out on paediatric diseases, as published in [5], the research showed that the Machine Learning Classifiers (MLCs) handles 101.6 million data points from a total of 1,362,559 pediatric patients. The hypothetic-deductive reasoning used by physicians and unearth associations are difficult to be found by the conventional statistical methods. D. S. Kermany et al. utilized the deep learning algorithm to identify medical diagnoses and treatable diseases [4]. The approach performance in classifying age-related macular degeneration and diabetic macular edema is comparable to human experts. The Machine Learning approach supports the decisions related to patient diseases and based on the imbalance classes proposed by [28] Previous studies defined the ML approach as the CCOA-RA used to overcome the imbalance negative and positive labels of the dataset.

Since its inception [29], the Extreme Gradient Boost (XGBoost) has been the favourite technique used to address the challenge of classification prediction in the medical area. By utilizing XGBoost, many researchers addressed various sub-topic in the health/medical field such as diagnoses [5], [8], [30]–[34], related diseases[4][6], [9], [34], [35], medical treatment [36], [37], patient status[12], [14], [16], [38], or event genomic [39].

A research conducted an experimental study using XGBoost Classifiers with some scenarios such as transformation, resampling, clustering, and ensemble learning to predict the diagnosis of second primary cancers (SPCs) [30] The resampling and clustering strategies were used to determine the best method used to identify some important risk factors associated with SPCs in patients with breast cancer. The combination of the XGBoost and Clustering analysis approach was also proposed by [8] to predict the hypertension-related symptoms from 531 hypertensive patients data in a hospital in Beijing. These combination techniques showed that there are significant differences in symptomatic entropy between patients with type II and type I hypertension. The experimental study of various classifiers techniques, such as XGBoost, was conducted by [32] to predict falls – non falls of Parkinson Diseases. Therefore, clinical, demographic, and neuroimaging data used in this study were obtained from Medical Centres, University of Michigan, and Tel Aviv Sourasky Medical Center. The research finding has a prediction accuracy value of 70 % - 80 %, which is used to provide a more reliable clinical outcome forecast of falls in Parkinson's patients. The superiority of XGBoost used to predict and diagnose Alzheimer-type dementia in the blood is divided into two categories, namely Alzheimer's Disease (AD) and cognitively normal (CN), according to [5]. In this study, experiments were conducted using some Classifier technique applied to 883 patient's data. The experiment finding shows that XGBoost gave the best performance. The high emergency diagnostics error rate is also commonly found in Urinary tract infection (UTI) due to clinical or physical symptoms. Machine learning based on the XGBoost technique has been demonstrated as the powerful tools used to overcome the challenge published in [32]. According to previous studies, the UTI prediction consists of six machine learning algorithms, with medical and social information. Therefore, the authors claim that XGBoost accurately diagnosed positive urine culture results.

III. Material and Method

A. Dataset

Data were collected from a total of 80,000 patient's laboratory test daily recorded by the Hospital Information System (HIS) in January 2019 as shown in Table I. The privacy records such as PATIENT_ID, PATIENT_NAME, and CLINICIAN were not presented in the table. Out of eight attributes, only four were used in the experiments, namely, TEST, AGE, SEX, RESULTS, and FLAG. The indicators of RESULT are High, Low, or Normal

TABLE I. Attributes and Value Sample of Raw Data

No.	Attribute	Value Sample
1.	Date	1/1/19
2.	Test	Number of Leucocyte
3.	Lab No.	19000001
4.	Age	15Y
5.	Sex	M
6.	Source	UGD
7.	Result	15.6
8.	Flag	H

B. Method

1) Research stages: The research stages globally consist of 3 blocks activities, namely data processing, modelling, and evaluation, as shown as the flow diagram in Fig. 1. The data processing stage comprises of collection and pre-processing activities. Some classifier techniques are used in the modelling stage, which is described in the next sections. In addition, the accuracy and AUC-ROC parameters were used to evaluate each technique performance.

Data were collected from 80.000 records of patients' laboratory test in HIS operated by a private hospital in Jakarta, Indonesia. All records were taken from transactions that were generated in January 2019.

This was followed by the processing step, which includes attribute removal, ignoring record with missing value, transforming rows to columns, value scaling normalization, and labelling. Some removable attributes are related to privacy such as patient and clinician's name as well as those attributes with no meaning to the study such is PATIENT_ID and LAB_ID. However, after row-column transformation and

ignoring the records with missing value, the dataset finally contains 4412 instants. Furthermore, labelling was manually performed in accordance with the patients' medical records. One of two instant data label is either 1 or 0, which represents inpatient and outpatient care. After the pre-processing step, each instant consists of 8 (eight) attributes, namely HAEMATOCRIT, HAEMOGLOBINS, ERYTHROCYTE, LEUCOCYTE, THROMBOCYTE, MCH, MCHC, MCV, AGE, SEX. The processed dataset is published in http://bit.ly/3kbK3Wn, and some of the instant samples, as presented in Table II.

2) Experimental scenario: Two experiments were carried out in accordance with the final data presentations. The first was conducted using the original value as shown in Table II, and labelled as Format 1:Lab Test Results, while the second scenario used the value transformed/coding of attributes as Format 2: Laboratory test result formatted attributes. In Format 1 the original value was the normalization scaling of 0 to 1 for all numeric value attributes. Meanwhile, the Format 2 dataset, used a rule regarding the laboratory test component obtained from some medical references[40], [41]. The rule is categorized into three levels, namely Low, Normal, and high, base on gender and age of the patient, as shown in Table III. For the attribute value transformation, 0, 1, and 2 were used to represent low, normal, and high, respectively.

In the first step of the modelling phase, to both of format data representation above we apply some classifier techniques such as the decision tree, Gaussian naïve Bayes, random forest, Adaboost, and XGBoost as shown in Fig. 1. The better format data representation is then used in the second step modelling phase, which also chooses the two best performances of techniques from the first step. In the first modelling step, the cross-validation scheme was chosen as the data testing splitting scenario since it is more representative compared to the random splitting process. Accuracy parameter was also used as the performance evaluation criteria, whereas in the second step, the AUC-ROC parameter was utilized.

The second step of the modelling phase was conducted by choosing the best two classifiers applied to the better format of data representation which gives the better performance. The hyperparameter optimization was performed on both classifiers to obtain best performance using the GridSearchCV which adapted from the skcit-learn library [42] In the second modelling we use the random splitting scenario of training – testing data selection, which the training data part is used in best hyperparameter searching and model training whereas the testing data part is for model validation (testing).

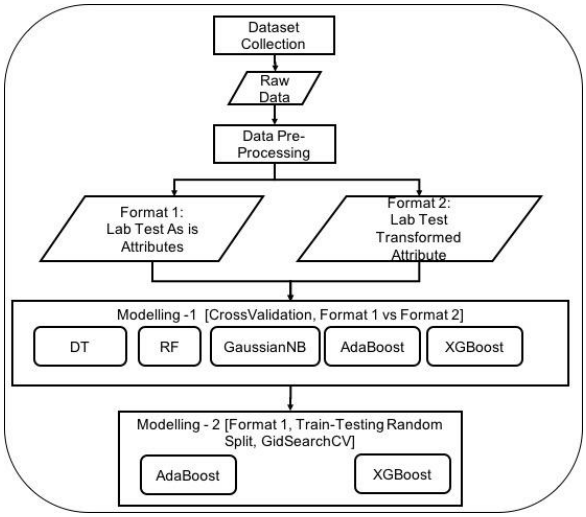


Fig. 1. Research Flow Diagram.

TABLE II. ATTRIBUTES AND VALUE SAMPLE OF RAW DATA

Attributes	Remark	Value Sample
HAEMATOCRIT	Continuous	35.1
HAEMOGLOBINS	Continuous	11.8
ERYTHROCYTE	Continuous	4.65
LEUCOCYTE	Continuous	6.3
THROMBOCYTE	Continuous	310
MCH	Continuous	25.4
MCHC	Continuous	33.6
MCV	Continuous	75.5
AGE	Continuous	12
SEX	Nominal - Binary	F

TABLE III. LABORATORY TEST RESULT FROM THE CATEGORIZATION RULE

	Low if less than*				**High if more than***			
Gender	***Male***		***Female***		***Male***		***Female***	
Age	***Adult***	***Infant***	***Adult***	***Infant***	***Adult***	***Infant***	***Adult***	***Infant***
HAEMATOCRIT	38,8	33	34,9	33	50	38	44,5	38
HAEMOGLOBINS	13		12		17		15	
ERYTHROCYTE	4,7	4,1	4,2	4,1	6,1	5,5	5,4	5,5
LEUCOCYTE	5	5	5	5	10	10	10	10
THROMBOCYTE	150	150	150	150	400	400	400	400
MCH	27	27	27	27	33	33	33	33
MCHC	32	32	32	32	37	37	37	37
MCV	80	80	80	80	96	96	96	96

**Normal If the Value Is In Between*

3) Evaluation criteria: This study used the commons criteria performance parameter in classification i.e. accuracy. The Accuracy computation is presented as formula (1) [43].

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

where:

TP : True Positive

TN : True Negative

FP : False Positive

FN : False Negative

The objective of the classification scenario is to achieve good quality performance of both classes. Therefore, the individual measures of both negative and positive classes are combined. The weakness of the accuracy parameters is none of them is adequate by itself. That's why we also use the other measure which is the common approach to unify those measures and produce an evaluation criterion such as the *Receiver Operating Characteristic (ROC)* graphic. ROC graphic represents the visualization of the trade-off between the benefits (*TPrate*) and costs (*FPrate*). The Area Under the ROC Curve (AUC) corresponds to the probability used in identifying the noise in the two stimuli [43]. Formula (2) presents the computation of AUC measure.

$$AUC = \frac{1+TPrate-FPrate}{2} \quad (2)$$

where

$$TPrate = \frac{TP}{TP+FN}T \quad (3)$$

$$FPrate = \frac{FP}{FP+TN} \quad (4)$$

TPrate is the percentage of positive instances correctly classified is positive class.

FPrate is the percentage of positive instances misclassified.

In the first step of the modelling phase, some classifier techniques as depicted in Fig. 1: the decision tree (DT), gaussian naïve Bayes (GaussianNB), random forest (RF), AdaBoost, and XGBoost are applied to the data representation. The better format data representation is then used in the second step modelling modeling step, which also chooses the two best performances of techniques from the first step. The cross-validation scheme was selected as the training and testing data testing splitting scenario because it is assumed the cross-validation is more representative compared to random splitting. Furthermore, the accuracy parameter is used as the performance evaluation criteria in the first step of the modelling phase, whereas in the second step, the model was evaluated by using the AUC-ROC parameter.

4) Overview of machine learning techniques: The five algorithms used to explore the experiments are presented in this section as follows:

a) Decision Tree: Decision tree (DT) is one of the techniques widely-used for classification purposes [44] The decision tree built is similar to a flowchart [45] and acts as a predictive model that contains a mapping between object values in the tree and the data attributes. The classifier is represented as the decision node of each instant data attributes, whereas the tree branches correspond to different prediction output. Each leaf node represents a possible output of the final presentation of the DT construction phase, which is expressed as the construction and pruning phases. In the construction phase, the overall planning of the DT main structure is completed, whereas in the pruning phase, a more precise pruning process is performed. The main advantage of DT compared to other methods is that it is very interpretable. In a certain field, such as healthcare, the interpretability is often preferred rather than the higher accuracy and relatively uninterruptable [44].

b) Random Forest: Random Forest (RF) is the combination of supervised and unsupervised learning algorithms capable of increasing the accuracy of machine learning classifiers [46]. As a multi-class classifier, it is resistant to noise, fast in training and classification, and has powerful classification capabilities [47]. Furthermore, the ensemble learning technique is based on a decision tree and widely used in various areas with almost ideal prediction [48]. Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi claim that RF is more robust than other famous models and have been utilized in many fields such as, computer visions and pattern recognition [49]. The common weakness in using RF is the processing needs more time when applied to large amounts of data because it has to build many tree models [50]. A large number of trees also require significant memory capacity [49]. The summarization of the main process to construct the RF is referred to [51].

c) Gaussian Naïve Bayes: The Gaussian Naïve Bayes (GaussianNB) algorithm is categorized as the supervised learning method [52] It is extended to real-valued attributes through a Gaussian distribution network. GaussianNB algorithm assumes that the probability of each attribute belonging to a given class value doesn't depend on all other attributes. When the value of the attribute is identified, the probability is called conditional. Data instances probability is computed by multiplying all conditional attributes. The prediction is formulated by computing each class instance and by selecting the highest probability class value [53].

d) AdaBoost: AdaBoost (Adaptive Boosting) is known as the most famous Boosting algorithms, as proposed by [54]. It is able to self-adjust the weak classifiers after learning, and it is sensitive to noise data and outliers. AdaBoost has the ability to avoid overfitting some tasks efficiently and boosts weak learners to converge and become stronger classifiers [55]. The improved performance is achieved with different types of algorithms with the outputs obtained from the combination of a weighted sum.

The sampling to train data is used to replace the random sampling, which places attention on training data that are difficult to process. Weak classifiers are combined by replacing the average voting with a weighted mechanism. The effectiveness of the integrated weak classifiers is guaranteed by equipping the weak classifiers that are effective with a higher

weight and equipping those that are ineffective with a lower weight. It is not necessary to determine the learning performance of the weak algorithm in advance since the classification accuracy of the integrated strong classifier depends on the accuracy of all weak classifiers. The Algorithm that interprets the main idea of AdaBoost can be referred in [56].

e) XGBoost: XGBoost is a scalable machine learning system for tree boosting, which is proposed as an alternative method for predicting a response variable using certain covariates [26]. The system is available as an open-source package and widely recognized in many challenges of machine learning and data mining. Based on a study conducted by [8], some of the advantages of XGBoost are as follows, supports linear classifier, regularization to control the model complexity, capability in using the second order of Taylor expansion, and some more.

IV. Results and Discussion

A. Data Analysis

Data analysis was conducted after some pre-processing methods such as attributes restructuring and instant with missing value removal applied to the dataset, which contains 4412 instants comprising of 1784 INPATIENT class and 2628 OUTPATIENT class. Some analysis exercises were conducted to determine the characteristics of the dataset. Univariate analysis is performed to determine the distribution pattern of each variable based on instant classes. Fig. 2 depicted the distribution histogram of HAEMATOCRIT features of each class, whereas Fig. 3 presents the MCH distribution pattern. These figures show that, in general, the distribution of HAEMOTOCRIT and MCH features are relatively similar between IN_PATIENT and OUT_PATIENT classes. The HAEMATOCRIT feature of both classes is normally distributed with the dominant value between 25 to 50 and the MCH of 25 to 32/33.

The remaining seven features also present the most similar pattern shown by these four features. The detailed patterns of these features are presented in this article's appendix. The high degree of feature distribution patterns is similar for both classes because the identification class belongs to a certain instant. Therefore, the classification task for the laboratory test result dataset is challenging.

Another analysis applied to the data is a multivariate correlation, as shown in Fig. 4. The first three features, namely HAEMATOCRIT, HAEMOGLOBINS, and ERYTHROCYTE, are highly correlated. The other high correlation is presented by MCH vs. MCV feature, and from the class point of view, the correlation is roughly the same for both classes.

B. First Modeling Step

In the first modeling step, six classifier techniques, namely, Decision Tree (DT), Random Forest (RF), Gaussian Naïve Bayes, Ada Boost, Ada Boost with DT as the basic learner, and XGBoost were evaluated and applied to two kinds of data representation. The Cross-Validation dataset splitting scenario was performed on the training and testing with KFold value of 10. The testing accuracy performance from the cross-validation experiments is shown in Tables IV and V. In general, for all techniques explored, the first format data represention presented better results compared to the second. This shows that the recoding value of the laboratory test made the data easier for human understanding but reduces accuracy.

For both formats used in data representation, XGBoost and AdaBoost obtained the best testing performance by average with slight differences. For the maximum value, the Random Forest had a testing accuracy of 0.7986, which is outperformed compared to AdaBoost with a maximum accuracy is 0.7964. XGBoost is still the best technique for the maximum results with a testing accuracy performance of 0.8009. Therefore, the second modeling step used both methods in the Format 1 data representation.

1) Second step of modelling: The GridsearchCV hyperparameter tuning of AdaBoost and XGBoost provided the performance results, as shown in Table VI. For both performance parameters, XGBoost achieved better results compared to Ada Boost. It also outperformed the training and testing steps with a different performance pattern. In the training stage, the difference between the accuracy and ROC-AUC performance parameter values between XGBoost and AdaBoost was 0.0671 and 0.0775, respectively, while at the testing phase, the difference was 0.0108 and 0.0166.

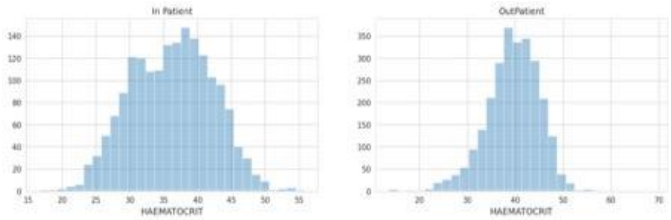


Fig. 2. Hematocrits Feature Distribution for Two Classes.

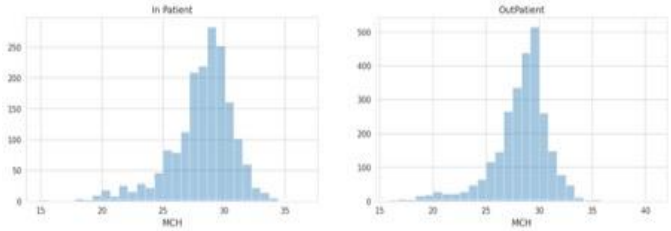


Fig. 3. MCH Feature Distribution for Two Classes.

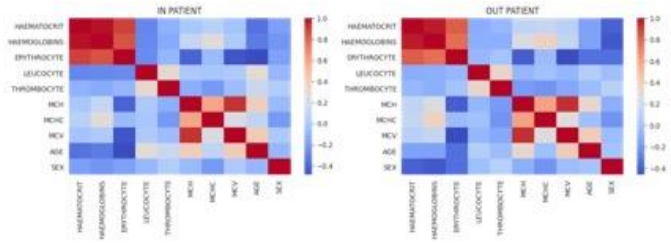


Fig. 4. Multivariate Analysis of Two Classes.

TABLE IV. TESTING ACCURACY PERFORMANCE OF FORMAT 1 DATA REPRESENTATION

	DT	RF	GaussianNB	AdaBoost	AdaBoost-DT	XGBoost
Max.	0.7014	0.7986	0.7392	0.7964	0.6825	0.8009
Min.	0.5646	0.6327	0.6131	0.6448	0.5692	0.6576
Ave.	0.6339	0.7158	0.6895	**0.7208**	0.6382	**0.7359**

TABLE V. TESTING ACCURACY PERFORMANCE OF FORMAT 2 DATA REPRESENTATION

	DT	RF	GaussianNB	AdaBoost	AdaBoost-DT	XGBoost
Max.	0.6281	0.6946	0.7308	0.7805	0.6719	0.7959
Min.	0.4751	0.5737	0.5737	0.4807	0.5510	0.6009
Ave.	0.5757	0.6335	0.6546	**0.6847**	0.6031	**0.6969**

TABLE VI. PERFORMANCE OF GRIDSEARCHCV HYPERPARAMETER TUNING

	Best Hyperparameter	Training		Testing	
		Accuracy	*ROC-AUC*	*Accuracy*	*ROC-AUC*
AdaBoost	n_estimators=230 learning_rate=0.1	0.7525	0.8145	0.7471	0.7936
XGBoost	learning_rate =0.1, n_estimators=1000, max_depth=5, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8, objective= 'binary:logistic', nthread=4, scale_pos_weight=1, seed=27)	**0.8196**	**0.8920**	**0.7579**	**0.8102**

Fig. 5 and Fig. 6 show the details of AdaBoost and XGBoost performance in ROC-AUC. The ROC curve shows that in any stage of specificity, XGBoost provides better AUC results compared to AdaBoost. Another information-insight shown by the curve is the different behaviour pattern of Training-Testing AUC for both classifiers. For XGBoost, Training – AUC is always on top of the Testing – AUC, whereas for AdaBoost in some part of curve Testing-AUC is the same.

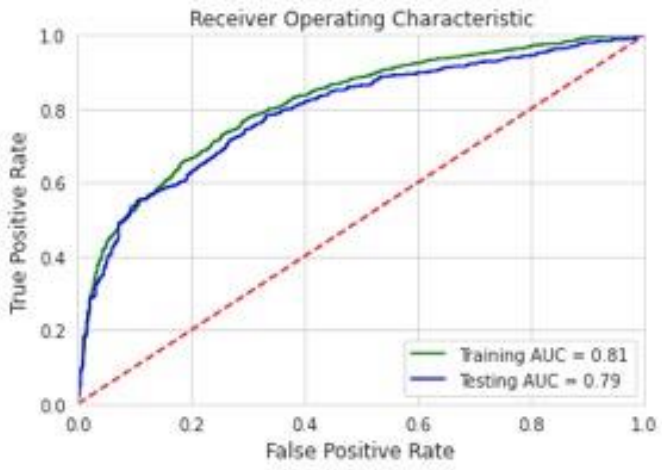


Fig. 5. AdaBoost ROC-AUC Graph.

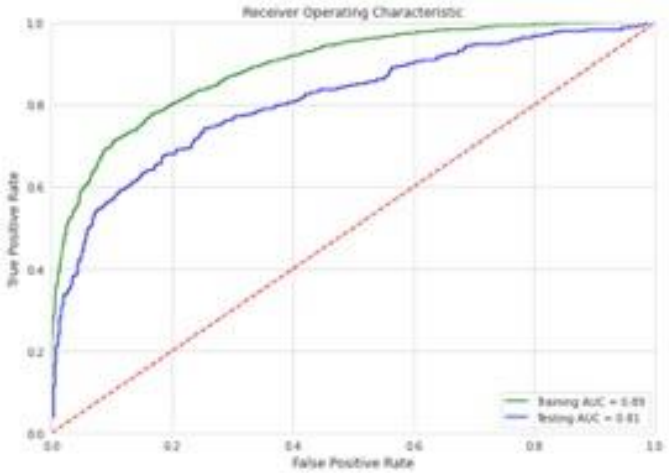


Fig. 6. XGBoost ROC-AUC Graph.

XGBoost classifier shows that THROMBOCYTE, AGE, and LEUCOCYTE are the top three factors of patient laboratory test results. Conversely, SEX is the least important attribute of patients, which contributes to the next treatment. The feature importance of patient data is shown in Fig. 7.

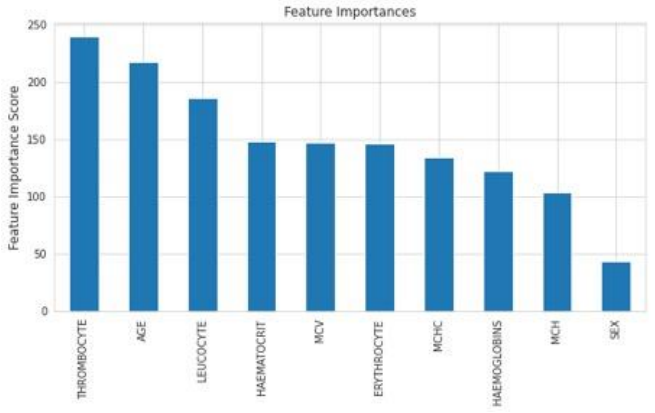


Fig. 7. Feature Importance by XGBoost Classifier.

V. Conclusion

In this research, the EHR dataset was collected from a private hospital located in Jakarta Indonesia to predict patient treatment recommendations. The work is one of limited computational study on a health-medical domain performed based on Indonesia local context. Based on data analysis, it can be concluded that the instant data characteristics belong to each class can be used to determine the next patient treatment. However, since the characteristics are quite similar, this condition makes it difficult to identify and classify challenges manually. The study shows that the XGBoost technique provides the best performance in predicting the next treatment to patients based on their laboratory test results. Another experimental result showed that THROMBOCYTE, AGE, and LEUCOCYTE are the most dominant feature in determining the class of a certain instant data.

The best testing accuracy achieved in the experiment is 0.7579. However, this is not acceptable in the health-medical field, which is related to human life. Therefore, many studies need to be carried out to overcome the obstacles. The limited information utilized as the input of machine learning techniques is one of the barriers addressed. Therefore, the use of additional patients' data such as their medical record history has the ability to improve the quality of the model. Future studies need to be conducted with easy access to patient information.

References

[1] M. Moscatelli et al., 'An infrastructure for precision medicine through analysis of big data', BMC Bioinformatics, vol. 19, no. Suppl 10, 2018. DOI: 10.1186/s12859-018-2300-5.

[2] A. Mavrogiorgou, A. Kiourtis, M. Touloupou, E. Kapassa, and D. Kyriazis, 'Internet of medical things (Iomt): Acquiring and transforming data into hl7 fhir through 5g network slicing', Emerg. Sci. J., vol. 3, no. 2, pp. 64–77, 2019.

[3] H.-Y. Liang et al., 'Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence', Nat. Med., vol. 25, 2019. DOI: 10.1038/s41591-018-0335-9.

[4] D. S. Kermany et al., 'Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning', Cell, vol. 172, no. 5, pp. 1122-1131.e9, 2018.

[5] D. Stamate et al., 'A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort', Alzheimer's Dement. Transl. Res. Clin. Interv., vol. 5, pp. 933–938, 2019.

[6] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, 'A data-driven approach to predicting diabetes and cardiovascular disease with machine learning', BMC Med. Inform. Decis. Mak., vol. 19, no. 1, pp. 1–15, 2019.

[7] M. Sharma and H. Aggarwal, 'Mobile based application for prediction of diabetes mellitus: FHIR Standard', Int. J. Eng. Technol., vol. 7, no. 2, pp. 117–120, 2018.

[8] W. Chang et al., 'Probability analysis of hypertension-related symptoms based on XGboost and clustering algorithm', Appl. Sci., vol. 9, no. 6, 2019. DOI: 10.3390/app9061215.

[9] X. Tian et al., 'Using machine learning algorithms to predict hepatitis B surface antigen seroclearance', Comput. Math. Methods Med., vol. 2019, 2019. DOI: 10.1155/2019/6915850.

[10] R. B. Lukmanto and E. Irwansyah, 'The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model', Procedia Comput. Sci., vol. 59, no. Iccsci, pp. 312–319, 2015.

[11] Suhaeri, N. M. Nawi, and M. Fathurahman, 'Early detection of Dengue disease using extreme learning machine', Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 5, pp. 2219–2224, 2018.

[12] C. A. Hu et al., 'Using a machine learning approach to predict mortality in critically ill influenza patients: A cross-sectional retrospective multicentre study in Taiwan', BMJ Open, vol. 10, no. 2, 2020. DOI: 10.1136/bmjopen-2019-033898.

[13] X. Hu et al., 'Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance', Eurasip J. Bioinforma. Syst. Biol., vol. 2016, no. 1, 2016. DOI: 10.1186/s13637-016-0049-6.C. Ye et al., 'Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm', Int. J. Med. Inform., vol. 137, no. February, 2020.

[14] S. Winiarti, H. Yuliansyah, and A. A. Purnama, 'Identification of Toddlers' nutritional status using data mining approach', Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 164–169, 2018.

[15] A. Shaham, G. Chodick, V. Shalev, and D. Yamin, 'Personal and social patterns predict influenza vaccination decision', BMC Public Health, vol. 20, no. 1, pp. 1–12, 2020.

[16] Kaneta Y, Tomida A, Tsuruo T, Nakamura Y, and Ohno R., 'Prediction of sensitivity to STI571 among chronic myeloid leukemia patients by genome-wide cDNA microarray analysis', Jpn J Cancer Res., vol. 93, no. 8, pp. 849–56., Aug. 2002.

[17] M. Sadikin, M. I. Fanany, and T. Basaruddin, 'A New Data Representation Based on Training Data Characteristics to Extract Drug Name Entity in Medical Text', vol. 2016, 2016.

[18] M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis, 'Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study.', J. Biomed. Inform., vol. 49, pp. 148–58, Jun. 2014.

[19] M. Sadikin, 'Mining relation extraction based on pattern learning approach', Indones. J. Electr. Eng. Comput. Sci., vol. 6, no. 1, 2017. DOI: 10.11591/ijeecs.v6.i1.pp50-57.

[20] L. Deléger, C. Grouin, and P. Zweigenbaum, 'Extracting medication information from French clinical texts', Stud. Health Technol. Inform., vol. 160, no. PART 1, pp. 949–953, 2010.

[21] Y.-F. Huang, H.-Y. Yeh, and V.-W. Soo, 'Network-based inferring drug-disease associations from chemical, genomic and phenotype data', 2012 IEEE Int. Conf. Bioinforma. Biomed., pp. 1–6, Oct. 2012.

[22] S. Saha and S. Rajasekaran, 'NRGC: A novel referential genome compression algorithm', Bioinformatics, vol. 32, no. 22, pp. 3405–3412, 2016.

[23] High-Throughput Variation Detection, and Genotyping Using Microarrays, David J. Cutler; Michael E. Zwick, and Minerva M. Carrasquillo,3, 'High-Throughput Variation Detection and Genotyping Using Microarrays', Genome Res, vol. 11, no. 11, pp. 1913–1925, 2001.

[24] P. R. M. Series, 'Contextual Factors: The importance of considering and reporting on the context in research on the patient centered medical home', 2013.

[25] F. Beenkens and P. Stolk, 'Context dependency of medical devices', 2010.

[26] D. Ramayanti et al., 'Tuberculosis Ontology Generation and Enrichment

Based Text Mining', in 2020 International Conference on Information Technology Systems and Innovation (ICITSI), 2020, pp. 429–434.

[27] H. Han, M. Huang, Y. Zhang, and J. Liu, 'Decision support system for medical diagnosis utilizing imbalanced clinical data', Appl. Sci., vol. 8, no. 9, 2018. DOI: 10.3390/app8091597.

[28] T. Chen and C. Guestrin, 'XGBoost: A scalable tree boosting system', Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 785–794, 2016.

[29] C. C. Chang and S. H. Chen, 'Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Breast Cancer Survivors', Front. Genet., vol. 10, no. September, pp. 1–6, 2019.

[30] R. A. Taylor, C. L. Moore, K. H. Cheung, and C. Brandt, 'Predicting urinary tract infections in the emergency department with machine learning', PLoS One, vol. 13, no. 3, pp. 1–15, 2018.

[31] C. Gao et al., 'Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease', Sci. Rep., vol. 8, no. 1, pp. 1–21, 2018.

[32] Y. Wang, Z. Du, W. R. Lawrence, Y. Huang, Y. Deng, and Y. Hao, 'Predicting hepatitis b virus infection based on health examination data of community population', Int. J. Environ. Res. Public Health, vol. 16, no. 23, 2019. DOI: 10.3390/ijerph16234842.

[33] S. V. Murty and R. Kiran Kumar, 'Accurate liver disease prediction with extreme gradient boosting', Int. J. Eng. Adv. Technol., vol. 8, no. 6, pp. 2288–2295, 2019.

[34] J. Taninaga et al., 'Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study', Sci. Rep., vol. 9, no. 1, pp. 1–9, 2019.

[35] X. Mo et al., 'Early and accurate prediction of clinical response to methotrexate treatment in juvenile idiopathic arthritis using machine learning', Front. Pharmacol., vol. 10, no. October, pp. 1–11, 2019.

[36] Z. Qiao, N. Sun, X. Li, E. Xia, S. Zhao, and Y. Qin, 'Using machine learning approaches for emergency room visit prediction based on electronic health record data', Stud. Health Technol. Inform., vol. 247, pp. 111–115, 2018.

[37] P. Kumar, A. Nestsiarovich, S. J. Nelson, B. Kerner, D. J. Perkins, and C. G. Lambert, 'Imputation and characterization of uncoded self-harm in major mental illness using machine learning', J. Am. Med. Inform. Assoc., vol. 27, no. 1, pp. 136–146, 2020.

[38] G. N. Dimitrakopoulos, A. G. Vrahatis, K. Sgarbas, and V. Plagianakos, 'Pathway analysis using xgboost classification in biomedical data', in ACM International Conference Proceeding Series, 2018, no. July 2018. DOI: 10.1145/3200947.3201029.

[39] A. Mulisin, 'Hematokrit : Nilai Normal, Rendah, dan Tinggi', 2020. [Online]. Available: https://www.honestdocs.id/hematokrit. [Accessed: 18-Dec-2019].

[40] Y. Capriyanti, 'Bahayakah Bila Jumlah Leukosit Berlebihan - Tanya Alodokter Tanya Dokter Diskusi Terbaru', 2020. [Online]. Available: https://www.alodokter.com/komunitas/topic/lekosit. [Accessed: 18-Dec-2019].

[41] L. Buitinck et al., 'API design for machine learning software: experiences from the scikit-learn project', 2013.

[42] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, 'An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics', Inf. Sci. (Ny)., vol. 250, pp. 113–141, 2013.

[43] D. Bertsimas and J. Dunn, 'Optimal classification trees', Mach. Learn., vol. 106, no. 7, pp. 1039–1082, 2017.

[44] A. Cubero-Fernandez, F. J. Rodriguez-Lozano, R. Villatoro, J. Olivares, and J. M. Palomares, 'Efficient pavement crack detection and classification', Eurasip J. Image Video Process., vol. 2017, no. 1, 2017. DOI: 10.1186/s13640-017-0187-0.

[45] M. Al-Qatf, Y. Lasheng, M. Al-Habib, and K. Al-Sabahi, 'Deep Learning Approach Combining Sparse Autoencoder with SVM for Network Intrusion Detection', IEEE Access, vol. 6, pp. 52843–52856, 2018.

[46] A. M. Joshi and S. Prabhune, 'Random forest: A hybrid implementation for sarcasm detection in public opinion mining', Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, pp. 5022–5025, 2019.

[47] Z. L. Li, H. X. Wang, Y. W. Zhang, and X. H. Zhao, 'Random forest–based feature selection and detection method for drunk driving recognition', Int. J. Distrib. Sens. Networks, vol. 16, no. 2, 2020. DOI: 10.1177/1550147720905234.

[48] Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi, 'Boosted random forest', IEICE Trans. Inf. Syst., vol. E98D, no. 9, pp. 1630–1636, 2015.

[49] N. Azizah, L. S. Riza, and Y. Wihardi, 'Implementation of random forest algorithm with parallel computing in R', J. Phys. Conf. Ser., vol. 1280, no. 2, 2019. DOI: 10.1088/1742-6596/1280/2/022028.

[50] W. Zhang, X. Zhao, and Z. Li, 'A Comprehensive Study of Smartphone-Based Indoor Activity Recognition via Xgboost', IEEE Access, vol. 7, pp. 80027–80042, 2019.

[51] M. C. Belavagi and B. Muniyal, 'Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection', in Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016), 2016, vol. 89, pp. 117–123.

[52] A. Al Bataineh, 'A comparative analysis of nonlinear machine learning algorithms for breast cancer detection', Int. J. Mach. Learn. Comput., vol. 9, no. 3, pp. 248–254, 2019.

[53] Y. Freund and R. E. Schapire, 'A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting', J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, 1997.

[54] S. J. Lee, T. Chen, L. Yu, and C. H. Lai, 'Image Classification Based on the Boost Convolutional Neural Network', IEEE Access, vol. 6, pp. 12755–12768, 2018.

[55] Z. Yang, L. Xu, Z. Cai, and Z. Xu, 'Re-scale AdaBoost for attack detection in collaborative filtering recommender systems', Knowledge-Based Syst., vol. 100, pp. 74–88, 2016.

Sentiment Analysis using Social and Topic Context for Suicide Prediction

E. Rajesh Kumar[1]*
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Vaddeswaram
Guntur 522502, Andhra Pradesh, India

K.V.S.N. Rama Rao[2]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Aziznagar, Moinabad (m), Hyderabad, Telangana

***Abstract*—In many fields, analysing large user-generated microblogs is very crucial and drawing many researchers to study. However, processing such short and noisy microblogs is very difficult and challenging. Most prior studies use only texts to find the polarity of sentiment and presume that microblog site is independent and distributed identically, ignoring networked data from microblogs. Consequently, not satisfied with performance motivated by emotional and sentimental sociological approaches. This paper proposes a new methodology that incorporates social and topic context to analyze sentiment on microblogs by introducing the meaning of structure similarity into social context. Unlike from previous research employing direct relations from user and by suggesting a new method to quantify structure similarity. In addition, to design the microblog semantic relation, topic context is introduced. The Laplacian matrix of these graph produced by these context combines social and topic context and Laplacian regularization is applied to the microblogging sentiment model. The Experimental results on the two datasets show that, the suggested model had reliably and substantially outperformed the baseline methods that is helpful for suicide prediction.**

***Keywords*—Social context; topic context; microblogging; Laplacian matrix; emotional and sentimental**

I. Introduction

Getting real user sentiment from huge collections of social media content created by users (e.g., microblogs) is a great challenge. Often, it is a great benefit and has a broad range of application opportunities for the sentiment of mining customers, such as business intelligence, recommendation system, customer management and relationship [1,2]. The role of automated sentimental study requires the system or machine to understand in deep of natural language [3], which has attained some results in formal analysis of sentiment related to text [4,5]. However, its output and performance are dropped when applied to microblogging sentimental analysis as it may consider text are independent and identically distributed. Microblogs are significantly shorter and have different type of expression compared to 'long formal text', e.g. 'it's so cooooooool!' and 'lol', aggravates the vocabulary sparsity problem. Conversely, social networking offers various kind of metadata, like user relation that can be controlled to boost the accuracy of sentimental analysis.

The study of effect on microblog sentiment analysis of other metadata a head of texts ('called social context') has recently exhibited more interest from researchers, such as applying direct user relationships to sentimental analysis models [6,7]. To support these approaches, two sociological theorems: emotional contagion [8], sentimental consistency [9] are used. As per social context, sentiment consistency is known as user context, it indicates that all post tends to have same sentiment label posted by same individual: Emotional contagion (EC) states that same opinion may appear to have for similar kind of people called friends context. While these studies were already exploited for sentimental analysis by considering the effects of direct relationship from users and ignoring the influence of indirect user relationship [6,7]. But social network connections are heterogeneous [10], so analysing sentiment analysis in microblogs using direct user relationship is not appropriate. For example, in Fig. 1, the blue dialog box signifies positive sentiment of text and red dialog box represents negative sentiment. Text reflected in black dialog box are the one to be categorized. From the user relation between Sam and John no direct relation exists but have common friends Alex and Joe. All users have different opinion about suicide, users may also be in a depressive mindset that may lead to suicide. Sam had posted a tweet related to suicide "I will never be unhappy", this sentence is a positive comment towards suicide. However, for a machine it is very difficult to detect the polarity for any sentence from its literal meaning. Further, by using direct user relationship among users to support sentimental analysis the text classification cannot be classified into classes as John's friend Alex and Joe has no comments on depression that results in a classifier error.

According to recent research, Indirect user relationships have recently been used into recommendation systems [11,34]. The principle of these works is that same preferences or behavioral patterns were found among similar users. However, based on small literature it studies the indirect user relationship in analysing sentiment. In same instance, homophily [12] has gained much more popularity with the growth of sociological theory. The principle is that interaction between similar individuals occurs at a higher rate than between dissimilar individuals [13], which has a significant impact on the creation of friendships. The data that flows through the network such as behaviour and culture appears to be limited. In addition, some indication of both negative and positive sentiment of homophily has found in social network [14].

Based on these research works; a new model is proposed to analyze microblog sentiment using user indirect relations by structure similarity (SS). This approach is by an assumption: similar user's opinions must be similar and tested this

*Corresponding Author

assumption experimentally. First, through mutual friend relationships, related users are found, and similarity matrix is established. Finding similar or related users through mutual friends [15,16] is a regular practice and new connections are generated by similarity [12]. In addition, the same opinion [15] may be shared by two users who may have a new link between them. The principle of this approach is to search for possible or potential relationship that could be friends among users and by considering them into model of sentiment analysis. Second, topic factors are created, and context matrix is of the subject is formed. On the same topic [13] the occurrence of homophily is highly significant and the context of the subject or topic in turn will better exploit the homophily theory. Finally, the context of user structure similarity and topic context are merged into a model of a graph, this graphs Laplacian matrix is used to evaluate microblogging sentiment. From Fig. 1, users Sam and John have common friends Alex and Joe. So, there can be several probabilities of being friends where they can express the same sentiment with specific probability by assumption. Therefore, Sam could also have negative comments on #suicide in response to Johns negative comments on #suicide, so the accuracy by relationship for sentimental analysis can be assured.

The main key contributions in this paper include:

- Proposing a technique for modelling homophily by applying structure similarity in social networks.
- The emergence of structural similarity as a replacement for user-direct relationships in the social context of microblogs.
- To model the semantic relationships between microblogs by introducing topic or subject context.
- Proposing a new sentimental analysis model for microblogs that integrates context of structure similarity, user context, topic, and text information context.
- Extensive evaluation of the proposed method using real-world datasets to recognize the working of proposed method.

II. Related Works

In this section, some related works about microblogging sentiment and sentimental analysis are reviewed.

A. Sentimental Analysis

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file "MSW_A4_format". Existing model consists of two major categories: machine learning and lexicon-based method. Method based on lexicon [14,15,32] generally uses SenticNet [16], SentiWordNet [17], to tag positive and negative labels for terms occurring in sentence, then by summarizing the complete sentence by tagged words the sentiment of the document can be judged. Methods based on lexicons that do not require polarity label datasets are unsupervised. These approaches however depend too much on lexicons and domain related due to polarity change of words from domain to domain. The methods of machine learning view sentiment analysis as an issue of text classification [18,19]. In these methods, using text features like unigram, bigram and word embedding are extracted and applied to different classification techniques such as NB, SVM and so on. Machine Learning (ML) techniques are supervised with polarity labels and typically requires lot of training data. Hence, Classification accuracy is related to size of the data.

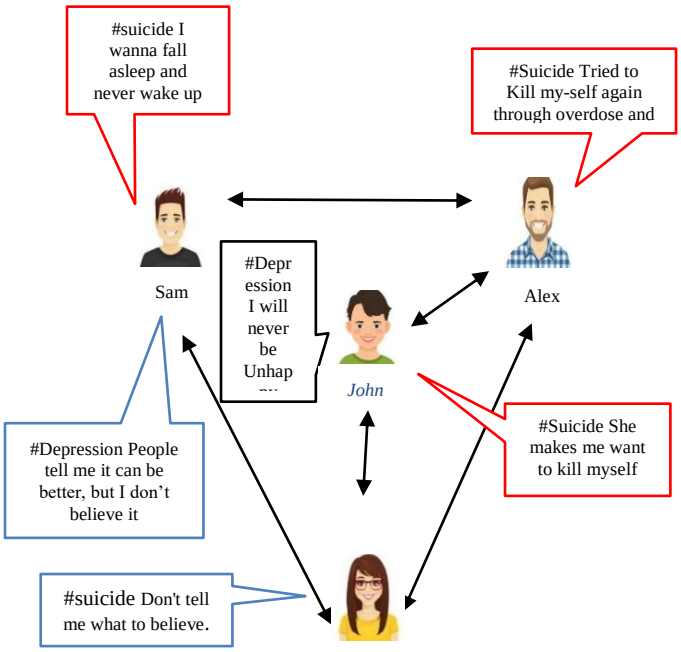


Fig. 1. User Direct Relationship.

B. Microblogging Sentimental Analysis

Over the years, microblog sentimental analysis is noisy, short, and become a hot research subject or topic [6,7,20] due to this problem many techniques are proposed to solve this issue. To analyze the opinion of tweets, [21] used emoticons features. In [22], repeated punctuations, generalized emoticons and words repeated were used to build a label propagation algorithm co-occurrence graph, this graph was used to identify the polarities of tweets sentiment. Using the relationship between emoticons and words, [23] used lexicon for feature sentiment extraction by developing a sentimental lexicon and for microblog analysis. All these above-mentioned strategies use only text information and ignores the additional information provided by microblog.

In recent years, there have been more and more studies on how to use user data to determine sentiment. [7,35] suggested a method to classify the sentiment of users on twitter using user '@' information and follow relationship. [24] brought user sentimental analysis to a specific subject or topic as a collaborative filtering problem, to predict user sentiment relationship between users were applied. Similarly, [25] has also manipulated the graph of user relation. In order to classify sentiment, entropy model with maximum result were used as labels and author applied label propagation approach. These working are method of classification of user topic level or user level sentiment, whereas the proposed model is microblogging level. In [6] author Hu et al., proposed a structure called SANT

('a sociological approach to noisy and short Text handling') that incorporates social context to characterize microblog sentiment. Based on [6], [26] included similarity of contents to the SANT system and suggested a semi-supervised model for tweet sentiment recognition. The author in [27] contended that context proposed by [6] was completely a content-based model, so for prediction level they suggested a framework for structured Microblog sentiment classification (SMSC). There also exists some works that have applied microblogs retrieval to user relationship [28]. All these approaches ignore user similarities and employ user direct relation.

III. Methodology

A. Dataset

In this paper, the experiment is performed on two different twitter datasets. These datasets were used for suicidal detection using different classification methods that included raw data with labelled sentiment.

Dataset1 is collected from microblogging websites consisting of trained dataset and test data set with labels positive and negative. It consists of five topics 'Evidence of suicide attempts', 'Suicide flippant reference', 'support or information', 'campaign or fight', 'suicide reporting', 'Condolence or memorial', 'None of these [33]. Dataset2 is created based on keywords [32]. Finally, users with friends are considered and remaining microblogs with no user friends are deleted.

B. Notations used

Uppercase letters such as M represents matrices, ***m*** indicates vector in bold, *m* represents scalar, M_{*i} is used to denote j^{th} column and M_{i*} denotes i^{th} row of the matrix. The matrix entry for row and column can be M_{ij}. Transpose of a matrix can be calculated by M^T. $||M||_F$ indicates Frobenius norm of matrix *M* and *tra(.)* for matrix traces.

The main objective of this paper is to construct a classifier W € Xm*p using the training matrix L € Xn*m (n- indicates feature and m represents number of microblogs) and labels the matrix Y € Xn*p (p indicates no. of polarities), classifier B € Xm*p is used to predict microblogs y that are unseen. Variable B indicates truth table, Ŷ=LC € Xm*c is used to represent matrix B truth table. Here, only binary classification is considered for sentiment, i.e., p=2. Consequently, the truth table is Bi*= [+1-1] for positive microblog and B*i = [-1+1] for negative microblog sentiment.

Consider an undirected graph G= (V, E) where, V indicates vertices and E represents edge. Mr represents microblog adjacency or relation matrix, Lm=Dm - Mr is Laplacian Matrix of G. Dm represents diagonal matrix and Dii is degree of ith vertex.

In Eq. (1) prediction feature is applied to identify microblog that are unseen. In Table II, variables, type, and their definitions are shown.

$$f(y)=\begin{cases}+1 & if \quad yC_{*1} > yC_{*2} \\ -1 & if \; yC_{*1} < yC_{*2} \\ +1 \quad or \quad -1 & randomly \quad if \; yW_{*1} = \quad yW_{*2}\end{cases} \tag{1}$$

TABLE I. Dataset Statistics

Emoticon	Dataset1	Dataset2
# of Twitter users	4562	5632
# of Twitter Tweets	147318	89141
# of Positive Tweets	53412	32765
# of Negative Tweets	78951	47697
Tweets avg. per user	32.21	15.84
Avg. friends per user	241.5	127.2

C. Microblog Content Modelling

For information related to text, standard least square method is used to satisfy the classification method. It aims to understand c-classifier and optimization problem after execution of (2) in terms of classification task in multiclass.

$$\min_y \frac{1}{2} \| LC - B \|^2_{R_m} \tag{2}$$

Unlike conventional text results, microblog leads to unigram sparse matrix due to noise and short in form. To deal this issue, sparse L1 regularization standard to seek for feature space by sparse reconstruction. To minimize the reconstruction error based on L1 norm, feature selection can be automatically implemented, and a sparse description can be obtained. To achieve a more stable model, L1 norm is implemented in the proposed model as shown in (3).

$$\min_y f(C;L;B) = \min_y \frac{1}{2} \| LC - B \|^2_{R_m} + \beta \| C \|_1 \tag{3}$$

Where, β is regulation weight.

D. Factors beside Text

In this section, various context is incorporated and integrated into a final model.

1) Integration of topic context: Hashtag is a mechanism that microblog services offer, using this service any user can insert information related to topics in microblogs. For instance, in a twitter tweets #symbols is used to tag tweet topic, let tweet "I wish to end my #life" represents tweets about "suicide thoughts". To express any kind of emotions, based on different topic people post in various microblogging site, in connection to same topic there can be same opinion by different users; different opinion for different topic, opinion of same topic with same person usually depend on each other. The importance of topic content is built to check whether more than one microblog text refer to same topic, to design connection with microblogs, it is better to include subject information into microblogging for sentimental analysis rather than text similarity. The microblog similarity value may be less by usage of text similarity leading to failure of sentiment efficiency. Using matrix-Mm in (4) "microblog-microblog matrix" for topic is obtained.

$$M_m = M_x + M_x{}^{M_t} \tag{4}$$

TABLE II. VARIABLE MEANING

Variable	Variable Meaning	Variable Type
M in Uppercase	Matrix representation	-
m in Lowercase	scalar representation	-
m in bold and lowercase	Vector representation	-
M_{*i}	Matrix-M (i^{th} column)	-
M_{i*}	Matrix-M (i^{th} row)	-
L	Feature matrix representation	X^{n*m}
B	Matrix truth table	X^{n*Sc}
$\hat{Y}$	Fitted label matrix	X^{n*Sc}
n	No. of features	*Int*
t	No. of training dataset	*Int*
Sc	Sentiment classification count	*Int*
x	No. of topic content	*Int*
C	Classifier	X^{t*Sc}
Y	Microblog feature vector	X^{t}
U_m	User Matrix	X^{r*n}
r	No. of users	*Int*
Ss	Structure similarity representation	X^{r*r}
M_t	Topic Matrix	X^{n*x}
M_m	Topic Microblog-Microblog Matrix	X^{n*n}
M_r	Relation Microblog-Microblog Matrix	X^{n*n}
D_m	Diagonal Matrix representation	X^{n*n}
L_m	Laplacian Matrix representation	X^{n*n}
R_m	User-user relation direct matrix	X^{r*r}
G	Graph representation	-
E	Edge representation	-
V	Vertices representation	-

Where, $Mt \in Xn*x$ is atopic matrix and $M_{t_{ij}}$ =1 if and only if (iff) i^{th} microblog is about jth topic information, $M_{m_{ij}}$ =1, if p_i and p_j microblog refers to same topic, D_m indicates diagonal matrix all assigned to zero.

2) Integration of user context: It is based on a theory called sentimental consistency. It recommends that tweet sentiment posted by same user in two microblog has greater probability than selection of random microblogs. Here, $M_r \in X^{n*n}$ indicates matrix sentiment consistency ('microblog-microblog').

To calculate $M_{r_{sc}}$ use (5), $Um \in X^{r*n}$ is a microblog matrix of user, where $U_{m_{ij}}$ =1, iff user-i post microblog-j and 'r' indicates no. of users.

$$M_{r_{sc}} = U_m{}^{M_t} + U_m \tag{5}$$

Where, $M_{r_{sc}}$ =1, iff p_i and p_j microblogs are posted by similar users.

3) Structure similarity content: This section is focused on sociological theory; emotional contagion, this indicates that sentiment posted by similar users in two microblogs has greater probability than selecting in microblogs randomly. In existing work, if sentiment posted by two different users in two microblogs with friends/followers related are connected, an approach is constructed to make two microblogs user sentiment to be closer as possible, it is known as friends' context. This is denoted by $M_{r_{ec}} = U_m{}^{M_t} * R_m * U_m$, where $R_m \in X^{r*r}$ indicates user matrix ('user-user') and $R_{m_{ij}}$ =1, iff there is a follower/followee relation among ith user and jth user. But existing work focused on direct user relationship and eliminating friends and follower's relationship.

As discussed in previous section, users can share the sentiment to the user "who is a friend of his/her friend", which is homophily communication. In this context, structure similarity is used to develop the EC sociological theory by considering friend relation. New relationship can be caused by common friends [29], example, if A and B have friend C in common, the friend probability will be increased between the users, this is known as "Triadic closure" [30].

The reality is A and B have friend relation to C that supports with confidence which is lacking with strangers at friendship creation is one of the causes for "Triadic closure". The second explanation is related on the motivation for C: it can minimize C's latent stress in two different relationship about bringing A and C closure.

In twitter with respect to three different users and three cases where users are closure and connected based on following two relations as shown in Fig. 2, 3 and 4. The incoming arrow to the user is pointed as followee and opposite user is follower. In case one as shown in Fig. 2, indicates the flow of communication between users, there can be a flow of opinion between Starc and Tony over Shane. In case two as shown in Fig. 3, two users tony and Starc with common followee Shane determining "friend of friend" relationship, if more followee's exist among two users, the construction of the relation can be made easier.

In case three as shown in Fig. 4, two users Tony and Starc with Shane as common follower. All three cases are indicating user similarity expressions, it implies possibility or relationship can be created between unconnected users. Due to this reason an undirected graph is taken for follower relationship.

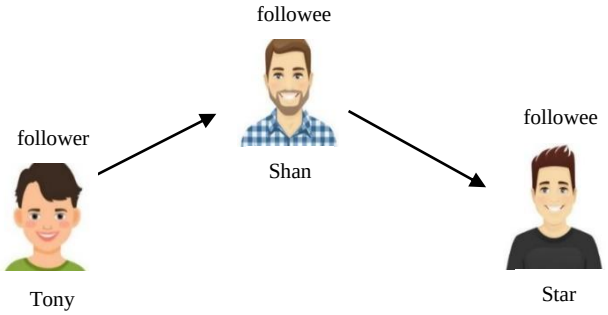


Fig. 2. Relation Type: Case One.

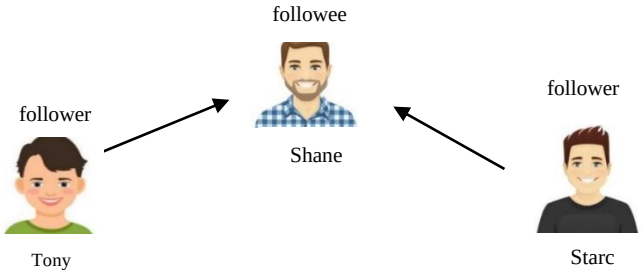


Fig. 3. Relation Type: Case Two.

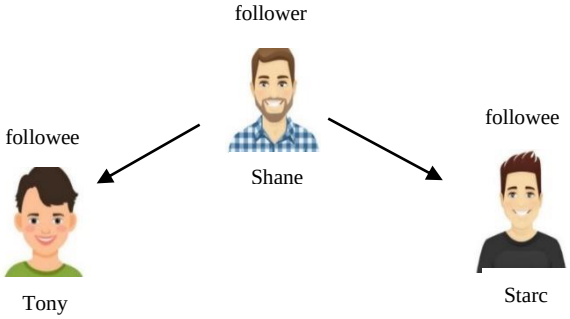


Fig. 4. Relation Type: Case Three.

Consider users v_i and v_j, Structure similarity can be analyzed by (6).

$$S_{s_{ij}} = sim(v_i, v_j) = |W_{ui} \cap W_{uj}| \tag{6}$$

Using similar friends between two user's Structural similarity (SS) can be determined. W_{v_i} indicates user V_i. $|W_{vi} \cap W_{v_j}|$ neighbours representing v_i and v_j number of friends in common. By including the condition as shown in Fig. 5, user Shane and Tony have common friends Joe and Starc. But in Fig. 6, user Tony has many friends, using (6) to calculate to SS for user Shane and user Tony in Fig. 6 will produce same SS value as obtained for Fig. 5. To manage this issue, all friends between two users are included to calculate SS.

$$s_{s_{ij}} = sim(v_i, v_j) = \begin{cases} \dfrac{|W_{v_i} \cap W_{v_j}|}{|W_{v_i} \cup W_{v_j}|} \\ \dfrac{|W_{v_i} \cap W_{v_j}|}{|W_{v_i} \cup W_{v_j}|} + 1 \end{cases} \tag{7}$$

Here, $W_{v_i} \cup W_{v_j}$ implies all set of friends (union) of both v_i and v_j users and $|W_{v_i} \cap W_{v_j}|$ indicates total users in the set. Once SS matrix Ss value is obtained, $M_{r_{ec}} \in$ C matrix can be calculated using (8).

$$M_{r_{ec}} = U_m^{M_t} * S_s * U_m \tag{8}$$

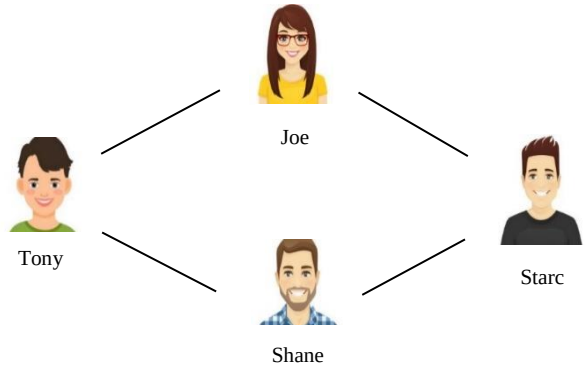


Fig. 5. Similarity Expression1.

4) Proposed model: Three types of context are incorporated in a framework. M_{r1} X^{n*n} represents both combination of SS context and user context that can be computed using (9). $M_{r2} \in X^{n*}$ indicates SS context, user context and subject or topic context, this can be calculated using (10).

$$M_{r_1} = M_{r_{sc}} + \beta * M_{r_{ec}} \tag{9}$$

$$M_{r_2} = (M_{r_{sc}} + \beta * M_{r_{sc}}) \circ M_m \tag{10}$$

Here, β=1, symbol denotes Hadamard product.

The main problem in existing system is microblogging text are noisy and short, there exist two kinds of problems. First problem is due to the large-scale nature of the vocabulary and dataset, this contributes to a high-dimension feature space. Second problem is the noisy and short text of data make data characterization extremely sparse. To avoid this problem sparse reconstruction is done as that change date robust to noise in terms of feature as discussed in Section 3.3.

The basic principle is to make two microblogs as identical as possible if they are posted by the similar users or two users similar to each other or same users based on sentiment consistency and emotional contagion to combine sentiment relation among microblogs in sentimental classification. This condition can be minimized by using (11).

$$= \min_C \sum_{l=1}^{c} \hat{Y}_{*l}^{T} (D^{M_t}{}_m - M_r) \hat{B}_{*l} \tag{11}$$

If only SS and user context is used then, $M_r = M_{r_1}$, $M_r = M_{r_2}$ for subject or topic content. Hence, final model combines social content and text information by using (12).

$$f(C; L; B) = \min_C \frac{1}{2} \| LC - B \|^2_{R_m} + \frac{\delta}{2} tri(C^{M_t} L^{M_t} L_m LC) + \beta \| C \|_1 \tag{12}$$

Where, δ indicates social content weight, β is regularization weight.

It is observed that (12) leads to non-smoothing optimization problem. It is reduced by convex smooth reformulation. To

solve this (12) is reformulated as shown in (13) which is a "constrained convex smooth optimization" problem.

$$\min_{C\in Z} L_m(C;L;B) = \frac{1}{2}\| LC-B \|^2_f + \frac{\delta}{r} tri(C^{M_t} L^{M_t} L_m LC) \quad (13)$$

Where, $z=\{C \| C \|_1 \le z\}$, differentiable part is $L_m(C;L;B)$, z -indicates non-differentiable part. L_{m_1} ball radius is z>=0, and one-to-one correspondence is present among z and β. For linear function $L_m(C;L;B)$, The smooth optimization problem is equivalently reformulated as proximal regularization [31] at C_x defines as $C_{x+1} = \arg\min_C G_{\gamma x C x}(C)$.

$$G_{\gamma x C x}(C) = L_m(C_x;L;B) + < \nabla L_m(C_x;L;B), C-C_x > + \frac{\gamma_x}{2}\| C-C_x \|^2_{R_m} \quad (14)$$

Algorithm1: Proposed Sentiment analysis using SS.

Proposed Sentiment analysis using SS **(PSASS)**

Input: L, B, C, δ, μ

Output: C

1. Randomly initialize C0
2. assign Ω=0,1, C1=C0, x=1
3. while till not convergence do
4. Calculate
5. Calculate
6. While condition is true do
7. Calculate
8. Calculate considering (16)
9. If then
10. Assign
11. Break
12. End if
13. Assign
14. End while
15. If x> max_iteration then
16. return
17. End if
18. Assign
19. Assign x=x+1
20. End while

Where, γx denotes size of step in x iteration. Therefore, the gradient of Laplacian matrix $L_m(C;L;B)$ respect to C can be evaluated using (15).

$$\nabla L_m(C;L;B) = L^x(LC-B) + \delta L^x L_m LC) \quad (15)$$

In considering β, Z constraint from (13) and (x+1)-th C can be calculated using (16).

$$(C_{x+1})_{j^*} = \begin{cases} \left(1 - \frac{\mu}{\lambda_x \| (U_m)_{j^*} \|}\right)(U_m)_{j^*}, if \| (U_m)_{j^*} \| \ge \frac{\mu}{\gamma_x} \\ 0, otherwise \end{cases} \quad (16)$$

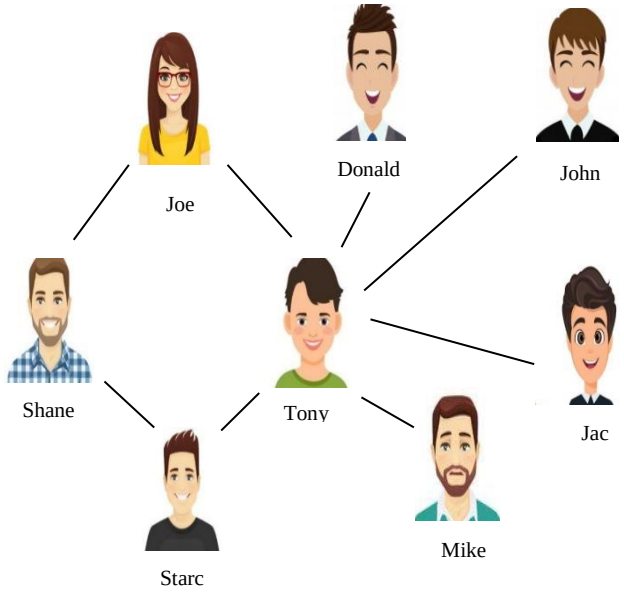


Fig. 6. Similarity Expression2.

Here, $U_{m_t} = C_x - \frac{1}{\gamma_x}\nabla L_m(C_x;L;B)$ For best convergence, optimization problem can be further accelerated and smoothened. Sequence C_x and V_x are utilized in the algorithm, C_x is sequence of estimated solution, V_x is combination affine of C_x and C_{x-1} is search point sequence. The final combination can be calculated by using (17).

$$V_x = C_x + \sigma_x(C_x - C_{x-1}) \quad (17)$$

Where, σ is the grouping coefficient. C_{x-1} suitable solution is calculated as "gradient" of V_x step through $G_{\gamma i V x}$. Finally, the algorithm for optimization is discussed as follows.

IV. SS AND SENTIMENT CORRELATION

The relation between sentiment label and friend's context are verified in [6,7]. A statistical analysis is done by illustrating how Sentiment labels in microblog and SS correlate. Consider G as an undirected graph where G=(V,E) is used to construct relation on two microblogs. Edge consisting of similar sentiment label is calculated by (18).

$$E = \frac{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(B_{i^*} = B_{j^*}, e_{ij} \in E)}{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(e_{ij} \in E)} \quad (18)$$

Where 1 is a function named indicator. The same calculation can be done in a weighted matrix using (19), the weights are regarded based on sentiment label. The same equation can be used for evaluating correlation among sentiment label in microblog and text similarity. In (19) I represent index, K indicates the weight of matrix G.

$$I = \frac{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(B_{i*} = B_{j*}, e_{ij} \in E).K_{ij}}{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(e_{ij} \in E).K_{ij}} \quad (19)$$

SS indicates two microblogging graphs built by SS, SS-topic represents two microblogging graph constructed SS and topic context. It is observed that ratio of SS and SS-Topic is greater between dataset1 and dataset2 as shown in Fig. 7. This possesses a positive relation between sentiment label and SS that helps to cover the study for exploring SS into microblogging sentimental analysis. Ratio of SS-topic method is said to be greater than SS model, it is because of homophily on similar topic and user may tend to have "same opinion on same topic".

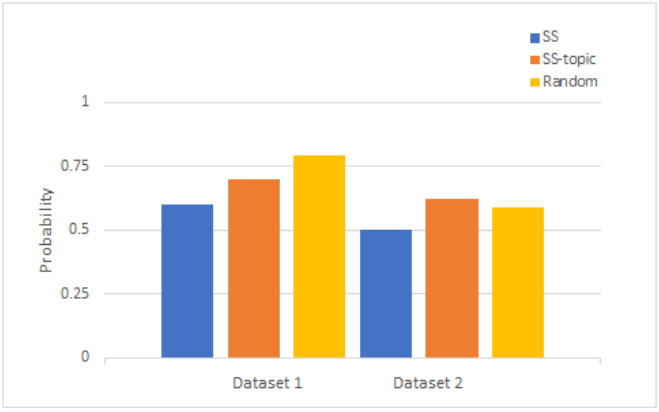


Fig. 7. SS-Conditioned Sentimental Polarity.

Thus, by including topic or subject content can explore heterogeneous relationship among microblogs.

V. Discussions

A. Advantage of Social Context

It is used to check the lifetime of various context whether they can improve in sentimental classification with respect to accuracy. 80% microblogging information are used to train the model. Here, "TC" denotes Text context, "UC" indicates user context and text. Likewise, "SSC" indicates text and SS context, "FC" implies Friends context and text. Accuracy is calculated as shown in Table III for the above-mentioned cases used by a metric, $accuracy = \frac{(TP+TN)}{(n)}$ where, n indicates number of negative and positive samples in the training dataset. TP is true positive, and TN is true negative labeled classes.

TABLE III. Context Performance

	FC	TC	UC	SSC
Dataset1	0.771	0.660	0.768	**0.799**
Dataset2	0.769	0.657	0.778	**0.788**

The following statements are determined from Table III.

- Using the "Social context", sentimental analysis on dataset1 and dataset2 can be improved in performance. Methods tested on Social content has better accuracy compared to text, it validates the utility of the context of the user, the context of friends and the context of structure similarity. The output of the social context illustrates that, in the microblogging network, emotional contagion and sentiment consistency (two theories) hold true.
- The user context performance is less compared to social contexts. This is mostly due to average friend's relationship usage post is more than the single user average post, contributing to more sentiment sparse consistency matrix. Example, according to Table I, every user in dataset1 has 32.21 average tweets with average friends 241.5.
- Approaches using SS context obtains best results in all context related to social text. SS the reason behind its better results than other context, it can get more data from friend's relationship than direct user relationship and weight on user who has more influence.

Structure similarity, which is the reason behind its better results than others, can get more data than direct relationships such as common friends and weights on whose effect is greater on users.

B. Performance Analysis and Comparison

Random sampling approach is used to check the accuracy of various methods with change in training dataset size. Methods are as follows:

1) Lasso: it is "Least absolute shrinkable and selection operator" is one of the regression analyses models which works on regularization and selection to improve accuracy of prediction it produces.

2) Least Square (LS): it is also one kind of approach in regression analysis, it is a statistical process for best fit to some set of information to be noticed. It is also used to predict related variable behaviour.

3) Support Vector Machine (SVM): it is one of a supervised ML model. They can categorize new text from the labelled training dataset for every group.

4) Naive bayes: it is one of a supervised ML model. The classifier of Naive Bayes assumes that the existence of a certain feature in a class is not linked to the presence of any other feature.

5) Logistic Regression (LR): it is also a statistical model used to design certain classes based on probability. It evaluates a dataset in which an outcome is calculated by individual variables.

6) SANT: Sentiment Analysis for Norwegian Text method proposed by [6] that combines two geological theories.

7) SMSC: Structured Microblog Sentiment Classification. All friends and user context are considered with equal priority.

8) SASS: Sentiment analysis based on SS is the proposed model to evaluate sentiment by using user context and SS. In this model, two initial positive parameters δ, μ are used. Assigning δ=0.0005 and μ=1 set by 'cross-validation'. Parameter μ denotes sparse regularization, parameter δ governs the information of social context contribution. For experimentation, from original or actual data testing dataset and training dataset are selected randomly. Percentage of the training dataset set is expressed by percentage and the remaining data for testing purpose. All the model's results are compared, and the observation are as follows:

- Models that use texts alone produced less result than methods using social context. Considered two samples and one tail T experiments are performed, and the results indicates that social context technique can boost the accuracy of sentimental classification with a significant level 0.01. In microblogging site, text data is very noisy, cynicism, and sarcasm are often used to convey user's negative emotions. Techniques like LS, NB, SVM and LR may not manage this scenario, by applying social context this problem can extended to some degree as the techniques take microblogs into account that are linked to perform a better result.
- SASS outperform SMSC and SANT and achieved better result on dataset1 and dataset2 with difference in size of training data significantly and consistently.

Compared with all models SASS has better performance with all different size of dataset with an accuracy of 0.799 for dataset1 and 0.792 for dataset2 as shown in Table IV. The performance is with respect to both dataset1 and dataset2 SANT and SMSC models had used friends and user context. But, in proposed model using SS it can explore relationship between microblogs intensely by using potential relationship between friends; every microblog possesses different impact to the suicidal sentiment compared to other microblogs but, in SMSC and SANT model all sentiment from microblogs has similar contribution to another microblogs.

SASS model is susceptible to change in size of the data for training. This shows that it is significant that "lot of labelling cost can be reduced" in spite labelling all training dataset manually.

9) Advantage of topic context: Topic context is introduced in proposed SASS model and SASS is compared with topic or subject context (SASS-T) varying with training dataset size from 60% to 90%. The result of classification is shown in Table V, from result it is observed that after incorporating topic context there is increase in accuracy of sentimental analysis in microblog compared to SASS model. Finally, SASS-T had produced better results 0.821 for dataset1 and 0.834 for dataset2 compared to all other models. The results signify the positive impact of applying topic or subject context in microblogging sentimental analysis to design the semantic relation among microblogs. The reason behind incorporating topic context is "the views of same person and similar kind of users on same topic usually remains consistent with each other that may help in prediction of suicide".

10)Parameter testing: Impact of two parameters δ, μ are tested for selection. These two parameters play a major role in managing the contribution of proposed model that might gain from SASS-T regularization constraints. The value of δ is assigned in some range {0,0.01,0.1,1,10 and 100} to study the impact on prediction accuracy. When SASS-T achieved better accuracy, the value of μ and δ are not same. So, the value of μ vary from {0, 1e-4, le-3,0.01,0.1,1} to study the impact of μ. When δ=μ=0, it reduces to SANT model, when δ>0, μ=0, it is SMSC model and when δ=0.005, μ>0, it is SASS model. So only suitable value of δ, μ can lead to a major improvement. Thus, proposed model has obtained better accuracy when δ=10, μ=0.1.

TABLE IV. PROPOSED AND BASELINE MODELS COMPARISION

	Training	**LS**	**LASSO**	**NB**	**LR**	**SVM**	**SANT**	**SMSC**	**SASS**
Dataset1 Without Topic content	60%	*0.644*	*0.693*	*0.759*	*0.724*	*0.718*	*0.770*	*0.762*	*0.776*
	70%	*0.622*	*0.665*	*0.768*	*0.731*	*0.725*	*0.763*	*0.771*	*0.772*
	80%	*0.612*	*0.695*	*0.754*	*0.733*	*0.731*	*0.759*	*0.767*	*0.779*
	90%	*0.612*	*0.721*	*0.745*	*0.729*	*0.745*	*0.769*	*0.761*	***0.799***
Dataset2 Without Topic content	60%	*0.653*	*0.677*	*0.702*	*0.718*	*0.709*	*0.701*	*0.723*	*0.733*
	70%	*0.659*	*0.692*	*0.705*	*0.716*	*0.718*	*0.717*	*0.722*	*0.744*
	80%	*0.662*	*0.671*	*0.691*	*0.709*	*0.723*	*0.730*	*0.739*	*0.774*
	90%	*0.645*	*0.740*	*0.681*	*0.722*	*0.731*	*0.749*	*0.755*	***0.792***

TABLE V. CLASSIFICATION-ACCURACY

	Training	**SASS**	**SASS-T**
Dataset1 With Topic content	60%	0.768	0.792
	70%	0.784	0.769
	80%	0.781	0.789
	90%	0.792	**0.821**
Dataset2 With Topic content	60%	0.736	0.754
	70%	0.750	0.756
	80%	0.771	0.772
	90%	0.792	**0.834**

VI. Conclusion

In this paper, a new technique is proposed to identify and facilitate sentiment classification as inspired by emotional contagion and sentiment consistency. Three types of context are considered in the proposed model: structure similarity, user context, and topic or subject context. structure similarity matrix and topic or subject context matrix are constructed, these contexts are added to the model using Laplacian matrix build by the contexts. experimental analysis showed that SS context produced better result compare to direct user relation. Thus, by incorporating topical context in SASS model aided in improving the outcome of sentimental classification compared to all other models with an accuracy 0.821 for dataset1 and 0.834 for dataset2. This result can be useful for suicide prediction among users based on the emotion of tweets posted by users that may help individual from attempting from suicide by informing to any NGO's.

VII. Future Work

The experiment is done on two different datasets with respect to topics and keywords with prediction results, further research is required based on different suicide risk factors that can be used for suicidal prediction and analysing the timeline tweets by examining retweets exhibiting suicidal contents with friends, followers, and tweeters.

Reference

[1] J, Mao H, Zeng X, Twitter mood predicts the stock market, Journal of Computational Science, vol.2, pp. 1-8, 2011.

[2] Cambria E, Mar, Affective computing and sentiment analysis. IEEE Intelligent Systems, vol.2, pp.102-107,2016.

[3] Cambria E, Schuller B, Xia Y, White B, New avenues in knowledge bases for natural language processing. Knowledge Based System, 108, pp.1-4, 2016.

[4] Fuji Ren, Ye Wu, IEEE transaction on affective computing, Predicting user opinions in Twitter with social and Topic content, vol.4, pp.412-424, 2013.

[5] Mei Q, Ling X, Wondra M, Su H, Zhai C, Topic sentiment mixture: modeling facets and opinions in weblogs, In Proceedings of the 16th international conference on World Wide Web-ACM, pp. 171-180, 2007.

[6] Hu X, Tang L, Tang J, Liu H, Exploiting social relations for sentiment analysis in microblogging, In Proceedings of the sixth ACM international conference on Web search and data mining, ACM, pp. 537-546, 2013.

[7] Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P, User-level sentiment analysis incorporating social networks, In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1397-1405, 2011.

[8] Hatfield E, Cacioppo JT, Rapson RL, Emotional contagion. Cambridge university press,1994.

[9] Abelson RP, Whatever became of consistency theory? Personality and Social Psychology Bulletin, 1983.

[10] Tang J, Hu X, Gao H, Liu H, Exploiting local and global social context for recommendation, In International Joint Conference on Artificial Intelligence, pp. 2712-2718, 2013.

[11] Tang J, Wang S, Hu X, Yin D, Bi Y, Chang Y, Liu H, Recommendation with social dimensions, The Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, pp. 251-257,2016.

[12] Mcpherson Miller and Smithlovin Lynn and Cook James M, BIRDS OF A FEATHER: Homophily in Social Networks. Review of Sociology vol.27(1), pp.415-444, 2001.

[13] Crimaldi Irene and Vicario Michela Del and Morrison Greg and Quattrociocchi Walter and Riccaboni Massimo, Homophily and Triadic Closure in Evolving Social Networks, arXiv: Social and Information Networks, 2015.

[14] Thelwall Mike, Emotion Homophily in Social Network Site Messages. First Monday, vol.15(4), 2010.

[15] Liang Y, Li Q, Incorporating interest preference and social proximity into collaborative filtering for folk recommendation, In Workshop on Social Web Search and Mining, 2011.

[16] Xie Yan Bo and Zhou Tao and Wang Bing Hong, Scale-free networks without growth, Physica A Statistical Mechanics & Its Applications, vol.387(7), pp.1683-1688, 2008.

[17] Liu J, Ji S, Ye J, Multi-task feature learning via efficient l 2, 1 -norm minimization. Conference on Uncertainty in Artificial Intelligence, AUAI Press, pp. 339-348, 2009.

[18] Ortigosa-HernaÂndez J, RodrõÂguez JD, Alzate L, Lucania M, Inza I, Lozano JA, Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. Neurocomputing 92, pp. 98-115, 2012.

[19] Wang Y, Huang M, Zhu X, Zhao L, Attention-based LSTM for aspect-level sentiment classification. Conference on Empirical Methods in Natural Language Processing. pp. 606-615, 2016.

[20] Pandarachalil R, Sendhilkumar S, Mahalakshmi G, Twitter sentiment analysis for large-scale data: an unsupervised approach, Cognitive Computation, vol. 7(2), pp. 254-262, 2015.

[21] Cheng C-H, Chen H-H, Sentimental text mining based on an additional features method for text classification. PLoS ONE, vol.14(6): e0217591, 2019.

[22] Cui A, Zhang M, Liu Y, Ma S, Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis, In Asia Information Retrieval Symposium. Springer, pp. 238-249, 2011.

[23] Kiritchenko S, Zhu X, Mohammad SM, Sentiment analysis of short informal texts, Journal of Artificial Intelligence Research, vol.50, pp. 723-762, 2014.

[24] Ren F, Wu Y, Predicting user-topic opinions in twitter with social and topical context, IEEE Transactions on Affective Computing, vol.4(4), pp.412-424, 2013.

[25] Speriosu M, Sudan N, Upadhyay S, Baldridge J, Twitter polarity classification with label propagation over lexical links and the follower graph, In Proceedings of the First workshop on Unsupervised Learning in NLP, Association for Computational Linguistics, pp. 53-63, 2011.

[26] Lu T-J, Semi-supervised microblog sentiment analysis using social relation and text similarity, In International Conference on Big Data and Smart Computing, IEEE, pp.194-201, 2015.

[27] Wu F, Huang Y, Song Y, Structured microblog sentiment classification via social context regularization, Neurocomputing 175, pp.599-609, 2016.

[28] Vosecky J, Leung KW, Ng W, Collaborative personalized Twitter search with topic-language models, international ACM SIGIR conference on research and development in information retrieval, pp. 53-62, 2014.

[29] Easley D, Kleinberg J, Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, 2010.

[30] Jackson MO, Rogers BW, Meeting Strangers and Friends of Friends: How Random Are Social Networks? American Economic Review, vol.97(3), pp.890-915, 2007.

[31] Kernighan BW, Lin S, An efficient heuristic procedure for partitioning graphs, The Bell System Technical Journal, vol.49(2), pp.291-307, 1970.

[32] E. Rajesh Kumar, K.V.S.N. Rama Rao, Soumya Ranjan Nayak & Ramesh Chandra, Suicidal ideation prediction in twitter data using machine learning techniques, Journal of Interdisciplinary Mathematics, vol.23:1, pp.117-125, 2020.

[33] Gualtiero B. Colombo, Pete Burnap, Andrei Hodorog, Jonathan Scourfield, Analysing the connectivity and communication of suicidal users on twitter, Computer Communications, pp.1-10, 2015.

A DNA Cryptographic Solution for Secured Image and Text Encryption

Bahubali Akiwate[1]
Department of Computer Science and Engineering
KLE College of Engineering and Technology, Chikodi, India

Latha Parthiban[2]
Department of Computer Science and Engineering
Pondicherry University, Pondicherry, India

***Abstract*—In recent days, DNA cryptography is gaining more popularity for providing better security to image and text data. This paper presents a DNA based cryptographic solution for image and textual information. Image encryption involves scrambling at pixel and bit levels based on hyperchaotic sequences. Both image and text encryption involves basic DNA encoding rules, key combination, and conversion of data into binary and other forms. This new DNA cryptographic approach adds more dynamicity and randomness, making the cipher and keys harder to break. The proposed image encryption technique presents better results for various parameters, like Image Histogram, Correlation co-efficient, Information Entropy, Number of Pixels Change Rate (NPCR), and Unified Average Changing Intensity (UACI), Key Space, and Sensitivity compared with existing approaches. Improved time and space complexity, random key generation for text encryption prove that DNA cryptography can be a better security solution for new applications.**

***Keywords*—DNA cryptography; image encryption; text encryption; DNA digital coding; DNA sequences**

I. Introduction

Security is often a crucial necessity for sensitive data transmission over communication networks. Various security techniques used to provide information privacy bring benefits to an organization or individual businesses [1]. There exist many benchmarks symmetric and asymmetric cryptographic algorithms such as Advanced Encryption Standard (AES), IDEA (International Data Encryption Algorithm), and RSA (Proposed by Rivest, Shamir, and Adleman) to provide security to text data. But the survey provides evidence that these algorithms are not suitable for image encryption [2, 3, 10, 24]. Image data characteristics like pixel correlation, bulk space, and high redundancy among pixel values make image encryption more challenging compared to text encryption [3,10]. Image Encryption plays a vital role in secured multimedia communication but the existing symmetric and asymmetric algorithms suffer from side-channel attack, Brute Force attack, Differential attack, and other statistical attacks [4]. There is a marked lack of better image cryptographic system. The proposed DNA based image Cryptosystem makes use of chaotic sequences to overcome existing limitations of symmetric and asymmetric cryptographic systems along with its confusion and diffusion properties [3]. To bring dynamicity, better storage, and time complexity, high parallelism, and low power consumption Adleman introduced DNA computing in 1994, which makes DNA cryptography the right choice for today's Internet applications [5]. DNA computing is still an area of interest for many researchers for its massively parallel processing capabilities and high resistance to brute force attacks [6]. The existing image encryption standards and mathematical models combined with DNA cryptography show defects in terms of CPU time, memory consumption, and battery usage [11]. The proposed DNA based approach for image encryption employs a chaotic sequence, which is deterministic and can produce a non-linear sequence [7]. It brings the advantages of unpredictability, pseudo randomness, and extremely sensitive to system control parameters and initial values [4, 7, 10, 13, 24, 27]. Also, Chaos systems can eventually return to the original state from the proceeded state [8-10]. The proposed approach involves a sequence of steps, such as the use of five-dimensional hyperchaotic sequences that produce a strong ciphered image, scrambling at the pixel level and bit level. The analysis of various parameters like Image Histogram, Correlation co-efficient, Information Entropy, NPCR, and UACI, Key Space, and Sensitivity shows that the proposed technique overcomes the limitations of the existing image encryption techniques. This paper also presents DNA based text encryption technique, which is based on the motivation of Kerckhoff's principle, which states that secrecy of transmitted message depends on key during decryption and not on an algorithm for encryption and decryption. At a high level, the algorithm is secure if the cryptanalyst is unable to deduce the key to obtain plaintext from the corresponding ciphertext [26]. This DNA based text encryption method uses the knowledge of random key generation to produce different sequences for the same input to achieve enhanced security performance. The proposed text encryption method outperforms the existing encryption techniques in terms of time and space complexities [11]. In this paper, Section 2 presents a preliminary study of the proposed approach; Section 3 covers image encryption in detail with result analysis. Section 4 discusses text encryption with two cases and time and space requirement analysis.

II. Preliminary Study

Traditional algorithms, including symmetric and asymmetric, are having many drawbacks concerning the exchange or use of a key. Compared to these, DNA cryptography can provide multifold security [45]. It provides an enriched security level [15, 16]. Conventional block cipher algorithms are not suitable for secured multimedia communication over public networks [2, 9, 24, 36, 37]. On the other hand, DNA cryptography is gaining more attention with a variation of chaos-based substitution permutation architecture [8, 16, 29, 30, 34, 36, 37]. It can run with lesser

memory and reduced computational overhead when compared with other standards like Elliptic Curve Cryptography, Packet wavelet, Fourier transform, Cellular automata, etc. [3, 8, 12]. The chaos method combined with DNA cryptography proved secure against a chosen-plaintext attack and differential attacks based on the previous studies [3, 17]. Boriga et al. proposed a 1D chaotic image encryption map that would be found weaker [23]. As there is a single variable is used, which makes easy prediction of initial values. Fidrich proposed use of 2D chaotic maps proved better diffusion properties. But, the use of a limited key space makes it easy to decode [18, 19]. Chen et al. proposed 3D chaotic maps proved better confusion and diffusion properties [20, 23]. Chen et al. presented a high complex 4D hyperchaotic system that brought many security advantages [38]. However, lower dimensional chaotic systems can be crackable as computer machines having limited precision. Understandably, only a portion of plaintext or ciphertext will help to get the key back [17]. Hence these are weaker against differential attacks [23].

The following are the demerits of previous works identified:

- Low dimensional chaos methods face difficulty in providing high security [30].
- Existing image encryption techniques work well only for a homogeneous image dataset, like medical or satellite images.
- There is no such practically used chaos-based DNA cryptography that exists worldwide in different application areas [4].

A. Our Contributions

The following are the major contributions of this proposed work:

- Capable of encrypting and retrieving highly sensitive images those are heterogeneous. The main features of these images are continuity, the large volume of data, and the strong association of adjacent pixels.
- During image encryption, pixel-level and bit-level permutation will render stronger cipher, which is difficult for an attacker to crack.
- The use of a 5D hyperchaotic system can ensure enhanced complexity and hence able to achieve improved security.
- Both image and text encryptions are possible over a single framework with better CPU latency.

III. DNA BASED IMAGE ENCRYPTION

The proposed image encryption technique performs a series of operations like scrambling using a 5D (five-dimensional) hyperchaotic method, XOR operations, and complementary rules to produce a solid ciphered image. [12,36]. It also includes scrambling at pixel and bit levels along with basic encoding rules and decomposition operations [17, 18]. Scrambling at the pixel level can be made according to a predefined concept [35,36]. DNA cryptography with a chaotic system depends on mathematical applications [19]. Edward Lorenz proposed the chaos theory in 1963 for the first time [35]. Confusion and Diffusion are two main properties of a chaotic system. With the confusion property, we can ensure the exchange of image pixel position randomly without affecting actual pixels. Diffusion mainly focuses on substituting one-pixel value with other pixel values by applying some mathematical operations over image pixels. Here only pixels will be permuted. Scrambling at bit level can make more difficulty in breaking the cipher by an attacker inducing reordering at bit levels. The proposed solution is a mixture of all these that can result in increased complexity for processes of encryption and decryption, making it harder for an attacker to crack.

A. DNA Digital Coding

DNA has four bases of Deoxyribo Nucleic Acid, namely Adenine (A), Thymine (T), Cytosine (C), and Genuine (G) [14, 19]. All the A and T bases complement each other according to the Watson - Crick Model. The bases C and G are mutually complementary [9,10,16,23]. Table I shows the DNA XOR operation between these bases [16, 21-23, 32].

TABLE I. XOR USE OF DNA SEQUENCES

XOR	**A**	**T**	**C**	**G**
A	A	T	C	G
T	T	A	G	C
C	C	G	A	T
G	G	C	T	A

Where A represents the binary value 00(Decimal Value 0), C represents the binary value 01(Decimal Value1), G represents binary value 10(Decimal Value 2) and T represents the binary value 11(Decimal Value 3) [16]. Every pixel represents a DNA sequence of length 4 in an 8-bit grayscale image. The proposed approach uses complementary rules for every character produced using DNA sequences. The complementary rule says about the base pairs. The bases, Adenine and Thymine can make one pair, and Cytosine and Genuine can make another pair [11-14, 21-23].

Suppose n_i be the set of bases A, T, C and G. Then complementary principle says the base string n_i of the encoding bases as follows:

$$n_i \neq C(n_i) \neq C(C(n_i)) \neq C(C(C(n_i)))$$

$$n_i = C(C\left(C(C(n_i))\right))$$

Where n_i and $C(n_i)$ are complementary and are base pairs. According to above statements, six complementary rules are as follows:

A → T,T → C,C → G,G → A

A → T,T → G,G → C,C → A

A → C,C → T,T → G,G → A

A → C,C → G,G → T,T → A

A → G,G → T,T → C,C → A

A → G,G → C,C → T,T → A

TABLE II. DNA ENCODING RULES

Rule	A	T	C	G
1	00	11	10	01
2	00	11	01	10
3	11	00	10	01
4	11	00	01	10
5	10	01	00	11
6	01	10	00	11
7	10	01	11	00
8	01	10	11	00

For each base of DNA sequence in terms of A, T, G, C their encrypted values will not remain same or equal [22].

DNA encoding and decoding operations are needed to map binary sequences into DNA bases and vice versa. Table II shows DNA encoding rules. There are 8 rules which satisfy the Watson–Crick complementary model [17, 19, 21, 24, 36, 37]. Here the selection of DNA bases such as A, T, C, and G can be made by following the DNA encoding rules. We consider two digits of the binary value for the mapping at a time.

Assume the original image P has scale matrix M X N. To bring out-diffusion property for an image P, Scrambling at pixel and bit levels are performed [15-17]. It permutes the bits of an image by considering the pixel values. Also it uses 5-D hyperchaotic system to obtain chaotic sequences as discussed below.

B. Hyperchaotic System

The proposed system uses five-dimensional (5-D) chaotic systems to enhance security and to bring increased complexity for image encryption [18]. It is covered below. A hyperchaotic 5-D system represented as the following equations (1):

$$c_1 = t(c_2 - c_1) + c_2c_3c_4$$

$$c_2 = u(c_1 + c_2) + c_5 - c_1c_3c_4$$

$$c_3 = -vc_2 - wc_3 - xc_4 + c_1c_2c_4$$

$$c_4 = -y_4 + c_1c_2c_3$$

$$c_5 = -z(c_1 + c_2) \quad (1)$$

Where t, u, v, w, x, y, z are system control parameters and c_1, c_2, c_3, c_4 and c_5 are system state variables. There are many existing approaches which used different algorithms to bring security for encryption. Since image data is highly sensitive, the chaotic sequence approach could be better to maintain security as there is the ability to overcome dependency on image pixels by confusion and diffusion properties [19-21].

C. Scrambling at Pixel Level

First, compute the initial values c_1, c_2, c_3, c_4 and c_5 of the 5-D hyper chaotic system as shown in equations (2) below:

$$c_1(1) = mod\left(\sum_{i=1}^{5} c_j^0, 1\right)$$

$$c_i(1) = mod\left(c_{i-1}(1) + c_j^0, 1\right) \text{ Where i=2, 3, 4, 5} \quad (2)$$

Where c^0_1, c^0_2 , c^0_3 , c^0_4 and c^0_5 are initial keys [22].

Producing chaotic series for image encryption might cause a transient effect that is impermanent and can have a sudden change of the state. Scrambling at the pixel and bit levels of an image can cause transient effect by internal or nearby values. So, there is a need for having several cycles to avoid such transient effect over hyperchaotic system N times, as shown in equation (3) below:

$$N = 200 + mod(\left(\left(\sum_{i=1}^{5} c_i^0\right) - \left|\sum_{i=1}^{5} c_i^0\right|\right) \times 10^{15}, 200) \quad (3)$$

By continuing the cycles or iterations over this chaotic system up to MN times, three chaotic sequences s_1, s_2 and s_3 are obtained.

Suppose the original plain image is P and its positions are (x, y). Let P' be the scrambled image of P and its positions are (x', y'). Then x' and y' can be calculated as below equations (4):

$$x' = x + mod(\left(abs\ (k_1(x) - \left|abs(k_1(x))\right|\right)\ \times\ 10^{15}, M - x)$$

$$y' = y + mod(\left(abs\ (k_2(y) - \left|abs(k_2(y))\right|\right)\ \times\ 10^{15}, N - y) \quad (4)$$

Where absolute values of x and y indicates the rounding of nearest integer values either lesser or equal to x and y.

By taking P as the input image, the scrambled image P' is obtained as follows (5):

$$P'(x, y) = P(x', y'), P(x', y') = P(x, y) \quad (5)$$

In above equation (5), Scrambled image P' is said to be positioned at (x, y), where, x=1,2,...M, y=1,2,…N.

D. Scrambling at Bit Level

Scrambled image P' is now converted into a sequence of one-dimensional values for P' from P'(1) to P'(MN) beginning with leftmost upper side to rightmost lower side of an image.

Now generate chaotic sequence s_3 as follows (6):

$$s_3'(r) = mod\ ((abs(s_3(r) - \left|abs(s_3(r))\right|)) \times 10^{15}, 8) \quad (6)$$

Where r = 1, 2,…MN, $s_3'(r) \in [0, 7]$

Scrambled image P' and decimal sequenced values of s_3' will be then transformed into binary sequences respectively. Then scrambled sequence C is obtained by having a circular shift over binary sequence P'(r) by considering the least bit of s_3' as following equation (7):

$$C(r) = circularshift[P'(r), LSB(s_3'(r)), s_3'(r)] \quad (7)$$

Finally, conversion from binary sequences into its equivalent decimal values can be required.

E. DNA Encryption

Calculate the initial values c_1' , c_2' , c_3' , c_4' and c_5' of 5D hyper chaotic system (1) as following equations (8):

$$c_1'(1) = mod\ (\sum_{j=1}^{6} c_j^0, 1)$$

$$c_i' = mod\ (c_{i-1}'(1) + c_i^0, 1) \quad (8)$$

Where i=2,3,4,5 and j=1,2,3,4,5,6

This system can have sudden changes over its state as there is image data. There is a need to iterate 5D hyperchaotic system N times, as shown below (9):

$$N' = 200 + mod\left((\sum_{i=1}^{6} c_i^0) - \left|\sum_{i=1}^{6} c_i^0\right|\right) \times 10^{15}, 200) \quad (9)$$

The chaotic sequences a_1, a_2, a_3 and a_4 are performed as follows (10-14):

$$a_1(i) = mod\left(\left(abs\left(a_1(i)\right) - |a_1(i))|\right) \times 10^{15}, 6\right) + 1 \quad (10)$$

$$a_2(i) = mod\left(\left(abs\left(a_2(i)\right) - |a_2(i))|\right) \times 10^{15}, 4\right) \quad (11)$$

$$a_3(i) = mod((abs\left(a_3(i)\right) - |a_3(i))| \times 10^{15}, 256) \quad (12)$$

$$a_4(i) = mod((abs\left(a_4(i)\right) - |a_4(i))| \times 10^{15}, 256) \quad (13)$$

Where $a_1 \in [1, 6]$, $a_2 \in [0, 3]$, $a_3 \in [0, 255]$, $a_4 \in [0, 255]$, i = 1, 2,.., 4MN.

Each C(r) and a_3(r) are expressed as below equations (14):

$$C(r) = \sum_{s=0}^{3} c_{4r-s}(r).4^s, c_{4r-s} \in \{0,1,2,3\}$$

$$a_3(r) = \sum_{s=0}^{3} d_{4r-s}(r).4^s, d_{4r-s} \in \{0,1,2,3\} \quad (14)$$

The sequences $\{c(i)\}^{4MN}_{i=1}$ and $\{d(i)\}^{4MN}_{i=1}$ can be constructed later. Where r = 1, 2, . . , MN.

Then Constructed sequences can be converted into DNA sequences. DNA sequence F(i) is obtained by using XOR operation as below (15):

$$F(i) = c'(i) \in d'(i) \quad (15)$$

Where i=1,2,…,4MN.

By using DNA replacement operation and complementary rules, get F'(i). F' can now be decoded to binary sequence G. Binary sequence G then translated into a decimal sequence.

Finally, an encrypted image R is obtained as below equations (16):

$$R(1) = a_4(1) \oplus mod(a_4(1) + D(1), 256) mod\left(\sum_{j=1}^{6} x_j^0 \times 10^{15}, 256\right)$$

$$R(i) = a_4(i) \in mod(a_4(i) + D(i), 256) \oplus R(i-1) \quad (16)$$

F. DNA Decryption

The decryption procedure is described as follows:

First, the chaotic sequences are generated and used (as shown above). Then decimal sequence D is obtained by the following equations (17):

$$D(1) = mod(a_4(1) \oplus R(1) \oplus mod\left(\sum_{j=1}^{6} x_j^0 \times 10^{15}, 256\right) - a_4(1), 256)$$

$$D(i) = mod\ (a_4(i) \oplus R(i) \oplus R(i-1) - a_4(i), 256) \quad (17)$$

By following DNA complementary rules and reverse replacement operations, now able to get F(i) as the DNA sequences. Now conversion from F(i) DNA sequence to C sequence is done by having bit-level scrambling. Then Convert C sequence into pixel-level scrambled image P'. Finally, the original image P is recovered from ciphered image P'.

Where D(i) is the value of the decimal sequence, a(i) is the value of the chaotic sequence, and R(i-1) is the value of the previous cipher pixel and R(i) is the value of the output cipher.

Fig. 1 shows the flowchart of the proposed image encryption system. The original image is converted into DNA sequences by the use of chaotic sequences obtained from the 5D hyperchaotic system and scrambling at pixel and bit levels. The DNA sequences are then converted into an encrypted image by following DNA XOR and Complementary rules [22, 41, 42]. Table III shows the images that were considered in the proposed approach.

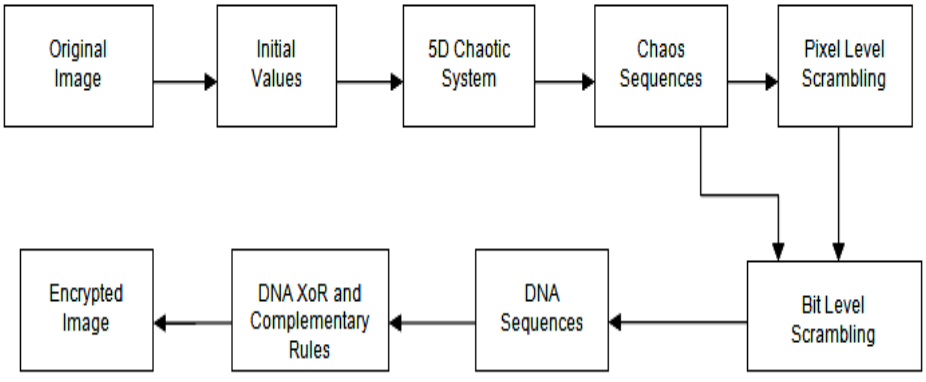


Fig. 1. Flowchart of Proposed Image Encryption System.

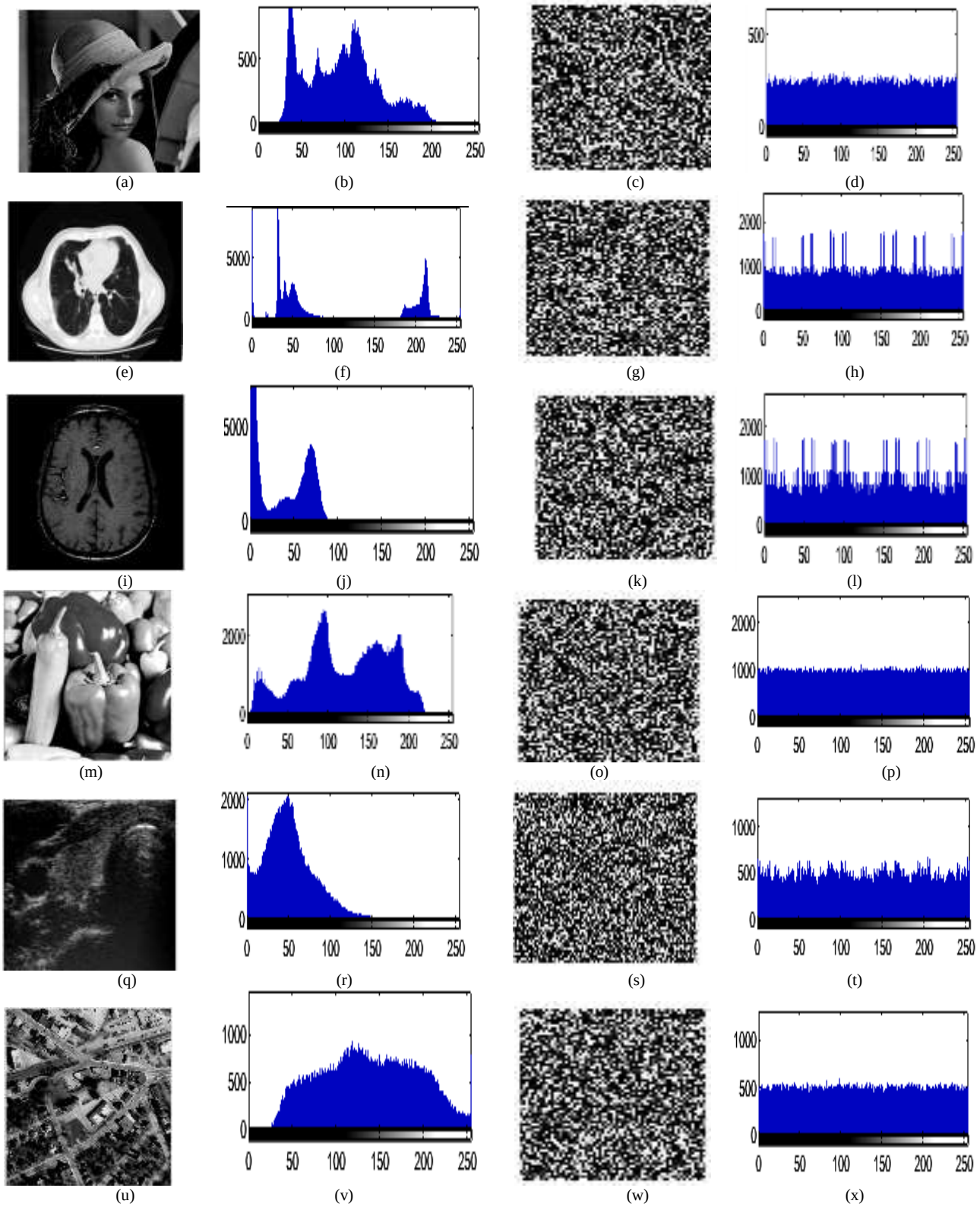


Fig. 2. (a) Original "Lena" Image (b) Initial Image Histogram "Lena" (c) Encrypted "Lena" Image (d) Final Image Histogram "Lena" (e) Original “ChestCT” Image (f) Initial Image Histogram "ChestCT " (g) Encrypted " ChestCT " Image (h) Final Image Histogram "ChestCT" (i) Original "MRI" Image (j) Initial Image Histogram "MRI" (k) Encrypted "MRI" Image (l) Final Image Histogram "MRI" (m) Original "Peppers" Image (n)) Initial Image Histogram " Peppers " (o) Encrypted " Peppers " Image (p) Final Image Histogram " Peppers " (q) Original "Ultrasound_Thyroid" Image (r) Initial Image Histogram " Ultrasound_Thyroid " (s) Encrypted " Ultrasound_Thyroid" Image (t) Final Image Histogram " Ultrasound_Thyroid" (u) Original " Aerial" Image (v) Initial Image Histogram "Aerial" (w) Encrypted "Aerial" Image (x) Final Image Histogram "Aerial".

TABLE III. IMAGES CONSIDERED

Image	Dimension (Width × Height)	Horizontal Resolution	Vertical Resolution	Bit Depth
Lena	256X256	96 dpi	96 dpi	8
Peppers	256X256	96 dpi	96 dpi	8
MRI	256X256	72 dpi	72 dpi	8
ChestCT	256X256	72 dpi	72 dpi	8
Ultrasound_Thyroid	256X256	72 dpi	72 dpi	8
Aerial	256X256	72 dpi	72 dpi	8

G. Image Encryption Results

1) Image histogram analysis: Image histogram represents the numbers of pixels that make an Image. Histogram Analysis helps us to understand the quality of image encryption [24]. A ciphered image histogram should have a uniform distribution [3,10,17,18,24,32]. Fig. 2 shows some centralized values for plain images(b,f,j,n,r,v) whereas there are more flat values for ciphered images(d,h,l,p,t,x) exists, which makes that the proposed system could withstand statistical attacks.

2) Key space and sensitivity metrics: The key space shows that all possible keys have been used [48]. Here chaotic sequences produced and used are combined along with precision value 10^{-15} to bring accurate refinement such as $c^0{}_1+ 10^{-15}$, $c^0{}_2+ 10^{-15}$, $c^0{}_3+ 10^{-15}$, $c^0{}_4+ 10^{-15}$, $c^0{}_5+ 10^{-15}$,…Hence, it leads to a larger key space around $(10^{15})^6=10^{90}=2^{298}$ which makes, this approach strong against brute force and dictionary attacks [16,17,31,40].

Key sensitivity refers to how much change in the key can impact to produce a ciphered image. Again this can be measured by parameters such as NPCR and UACI discussed above [18, 19]. A good approach is always sensitive, even a small change in the key to bringing out more diffusion or permutation in an image [39]. Hence the proposed approach is said to be resistive against differential and statistical attacks.

3) Correlation co-efficient analysis: Correlation coefficient values indicate the relationship between the pixels which are adjacent to each other [16, 17]. Smaller the values of correlation co-efficient show the greater security against attacks as resisting ability against them [5, 16, 27, 31]. Table IV lists the correlation coefficient values of six different images with diagonal, horizontal, and vertical values.

The proposed system could able to produce smaller co-efficient values either in diagonal, horizontal, or vertical directions when compared to existing approaches listed in the following Table IV for images Lena, Aerial, and Peppers. The correlation coefficient c_{pq} is computed as follows (18) [1,18,21]:

$$c_{pq} = \frac{cov(p,q)}{\sqrt{D(p)D(q)}} \quad (18)$$

TABLE IV. CORRELATION CO-EFFICIENT VALUES

Images	Diagonal	orizontal	Vertical
Lena	0.001809	0.001851	0.002981
ChestCT	0.002234	0.001511	-0.001400
MRI	-0.000581	-0.000074	0.000328
Peppers	0.000790	0.000901	0.004058
Ultrasound_ Thyroid	-0.000291	-0.000806	0.000476
Aerial	0.002243	-0.000986	0.003369
Lena[16]	0.0110	3.4459e-004	−0.0064
Lena[22]	0.0010	0.0068	−0.0054
Lena[29]	0.008006	0.011816	−0.017311
Aerial[16]	0.0110	−0.0109	−0.0211
Peppers[18]	−0.0107	−0.0010	−0.0292
Peppers[19]	0.0020185	−0.0013436	0.0066809
Peppers[29]	−0.009679	−0.015974	0.035035

Where p and q are adjacent pixels. $cov\,(p,q)$ is the covariance between two pixels p and q. It is given as follows (19):

$$cov(p,q) = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(p_i - E(p)\right)(q_i - E(q) \quad (19)$$

Where

$$E(p) = \left(\frac{1}{N}\right)\sum_{i=1}^{N} p_i$$

$$D(p) = \left(\frac{1}{N}\right)\sum_{i=1}^{N} (p_i - E(p))^2$$

4) Information entropy: Shannon introduced Information Entropy in the year 1948, which describes how much information is provided by an image. It helps us to understand the uncertainty or randomness level of an image [3, 16, 29, 43]. Uncertainty level before and after image encryption is measured. Reduced value of entropy indicates lesser the information provided by the encrypted image [18, 27, 31, 32]. The entropy of information should be 8 for an 8 bit image [30]. And it must be having the range 0 through 8. Table V shows Information entropy values for six different images considered. The proposed approach can produce a higher value of information entropy for an image sample Peppers when compared to some existing approaches.

Let n be the source of information. So the entropy of information can be measured as follows (20):

$$H(n) = -\sum_{i=1}^{L} p(n_i) log_2 p\,(n_i) \quad (20)$$

Where $p(n_i)$ is the probability of appearance of variable n_i and L indicates the length of an information in terms of total number of pixel variables.

TABLE V. INFORMATION ENTROPY VALUES

Images	Entropy
Lena	7.996402
ChestCT	7.951614
MRI	7.939766
Peppers	7.999182
Ultrasound_Thyroid	7.989553
Aerial	7.998492
Peppers[17]	7.9973520
Peppers[19]	7.9963
Peppers[22]	7.9967
Peppers[29]	7.997275

5) Analysis of NPCR and UACI metrics: The NPCR (Number of Pixels Change Rate) represents the number of different pixels in two images. In other words, NPCR helps us to understand the effect of change of single-pixel over an image. The UACI (Unified Average Changing Intensity) represents the difference in average pixel values of intensity between two images [1-5, 16-18,27, 29, 30-32]. Here the original image and an encrypted image for computations are considered. Ciphered image will significantly change if there is a tiny change in the pixel of a plain image.

Suppose C_1, C_2 are two ciphered images. Assume that these ciphered images are having a single pixel difference with their corresponding plain images. Let $C_1(i,j)$ and $C_2(i,j)$ at row i and column j respectively are two gray-level representations of the ciphered images C_1 and C_2.

The NPCR is obtained as follows (21):

$$NPCR = \left(\sum_{i=1}^{W}\sum_{j=1}^{H} R(i,j) \times 100\%\right)/(W \times H) \quad (21)$$

Where

$$R(i,j) = \begin{cases} 1 & C_1(i,j) \neq C_2(i,j) \\ 0 & C_2(i,j) = C_2(i,j) \end{cases}$$

The UACI is calculated as follows (22):

$$UACI = \left(\sum_{i=1}^{W}\sum_{j=1}^{H} \frac{C_1(i,j) - C_2(i,j)}{255}\right) \times 100\%)/(W \times H) \quad (22)$$

Where W and H correspond to the width and height of the image. Table VI shows the values of NPCR and UACI of various images. It is found that when compared with existing approaches, the proposed approach can obtain a higher UACI value for the encrypted Lena picture. The NPCR value for encrypted Lena image is almost nearer to existing approaches. A higher value of the NPCR and UACI means the system is safer against differential attacks.

TABLE VI. NPCR AND UACI VALUES

Images	NPCR	UACI
Lena	99.6279	49.7571
ChestCT	11.3342	2.8501
MRI	99.2271	24.9121
Peppers	99.2271	24.8966
Ultrasound_Thyroid	11.3074	5.7084
Aerial	99.2344	26.1513
Lena[12]	99.7570	39.12
Lena[16]	99.6067	33.4951
Lena[17]	99.6135	30.9255
Lena[19]	99.5892	33.4358
Lena[22]	99.61	33.46
Lena[29]	99.608337	33.431251

IV. DNA BASED TEXT ENCRYPTION

Symmetric algorithms are quicker in performing computations. One problem with these algorithms is more security breaches since single key usage. It means if an intruder receives the single key shared over a public channel, it can hack the entire network [9, 28]. Public key cryptographic algorithms, on the other hand, have proved adequate security for the systems. But here, more time is required to perform computations. There are some recent works in which the concept of DNA cryptography is combined with traditional algorithms such as AES, RSA, and ECC have provided better security for text-related encryption and transmission in the current computing age [20,23,33]. Due to its uniqueness, randomness, increased storage capabilities, high parallelism DNA Computing is gaining more popularity in cloud computing, Ubiquitous computing areas [15].

This section discusses, along with the study of its performance, DNA-based text encryption and decryption processes.

A. DNA based Text Encryption/Decryption

The text encryption/decryption uses DNA encoding rules including a single-point fusion, mutation, and complementary rules. Single-point cross over is the one where two bases are merged in order to build other bases. Mutation means modification in a DNA sequence by some means [25]. Here in every encryption, a random key will be generated and is used to encrypt data. The same is used during decryption to decrypt the ciphertext. It is important to generate random keys that are to be used to preserve the dynamicity of the proposed work [26, 28]. It means different transactions use different keys to produce different ciphertexts. It makes a more difficult cipher to break by an attacker. The Algorithm below shows the step-by-step procedure for the text encryption method of the proposed model. The decryption technique is precisely the reverse of an encryption process [44, 46].

Algorithm: Text Encryption Process

1. Start:
2. Read the plain text;
3. Generate Random Key
   Display Key length in bits and its value
4. Initialize round_no and decryption key
5. Invoke generate_preprocessing_tables ()
   Conversion from two bits to DNA bases
   Conversion from DNA bases to bits
6. Invoke generate_mutation_tables ()
   Conversion from four bits to two DNA bases
   Conversion from two DNA bases to four bits
7. Start Encryption Time
8. Invoke dnaChaosSecFuncE ()
   Define the number of rounds
   Get binarized data of text
   Convert binarized data into DNA sequences
   Display Initial DNA sequence
9. While (number of rounds > 0)
   Conversion from DNA bases into bits and encrypt the data using Key
10. Invoke crossover ()
   Invoke single_point_crossover ()
   Invoke rotate_crossover ()
11. Invoke mutation()
   Follow DNA complementary rules
12. Invoke reverse_reshape()
   Return reverse_reshape
13. End While ().
14. Display Final DNA Sequence
15. End Encryption Time
16. Display total execution time for encryption
17. End

In the above algorithm two bits values used are '00', '01', '10', '11'. Initial DNA bases are 'A', 'C', 'G', 'T'. Four bits values are '0000', '0001', '0010', '0011', '0100', '0101', '0110', '0111', '1000', '1001', '1010', '1011', '1100', '1101', '1110', '1111'. Two DNA bases used are 'TA', 'TC', 'TG', 'TT', 'GA', 'GC', 'GG', 'GT', 'CA', 'CC', 'CG', 'CT', 'AA', 'AC', 'AG', 'AT'.

The following content shows the operations that are taken place during the decryption of encrypted text.

<reshape>4<reshape><crossover><type>both<type><rotate><rotation_offset>2<rotation_offset><rotation_types>right|left|right|right|right|<rotation_types><rotate><single_point>2|3|<single_point><crossover><mutation><mutation_table>{'A':'C','C':'A','T':'G','G':'T'}<mutation_table><chromosome><complement_mutation>(0,5)<complement_mutation><alter_mutation>(1,3)<alter_mutation><chromosome><chromosome><complement_mutation>(5,6)<complement_mutation><alter_mutation>(3,3)<alter_mutation><chromosome><chromosome><complement_mutation>(2,5)<complement_mutation><alter_mutation>(3,3)<alter_mutation><chromosome><chromosome><complement_mutation>(3,6)<complement_mutation><alter_mutation>(0,1)<alter_mutation><chromosome><chromosome><complement_mutation>(1,4)<complement_mutation><alter_mutation>(2,3)<alter_mutation><chromosome><mutation><round><round><reshape>2<reshape><crossover><type>rotate_crossover<type><rotate><rotation_offset>2<rotation_offset><rotation_types>right|left|right|right|left|left|right|left|right|left|<rotation_types><rotate><crossover><mutation><mutation_table>{'A':'T','T':'A','C':'G','G':'C'}<mutation_table><chromosome><complement_mutation>(0,3)<complement_mutation><alter_mutation>(0,0)<alter_mutation><chromosome><chromosome><complement_mutation>(3,3)<complement_mutation><alter_mutation>(1,1)<alter_mutation><chromosome><chromosome><complement_mutation>(1,1)<complement_mutation><alter_mutation>(0,0)<alter_mutation><chromosome><chromosome><complement_mutation>(2,3)<complement_mutation><alter_mutation>(0,1)<alter_mutation><chromosome><chromosome><complement_mutation>(2,)<complement_mutation><alter_mutation>(0,1)<alter_mutation><chromosome><chromosome><complement_mutation>(3,3)<complement_mutation><alter_mutation>(1,1)<alter_mutation><chromosome><chromosome><complement_mutation>(0,0)<complement_mutation><alter_mutation>(1,1)<alter_mutation><chromosome><chromosome><complement_mutation>(0,0)<complement_mutation><alter_mutation>(1,1)<alter_mutation><chromosome><chromosome><complement_mutation>(1,2)<complement_mutation><alter_mutation>(0,0)<alter_mutation><chromosome><chromosome><complement_mutation>(1,3)<complement_mutation><alter_mutation>(0,0)<alter_mutation><chromosome><mutation><round><round><reshape>4<reshape><crossover><type>single_point_crossover<type><single_point>0|3|<single_point><crossover><mutation><mutation_table>{'A':'G','G':'A','T':'C','C':'T'}<mutation_table><chromosome><complement_mutation>(2,6)<complement_mutation><alter_mutation>(3,3)<alter_mutation><chromosome><chromosome><complement_mutation>(6,6)<complement_mutation><alter_mutation>(0,0)<alter_mutation><chromosome><chromosome><complement_mutation>(6,7)<complement_mutation><alter_mutation>(0,1)<alter_mutation><chromosome><chromosome><complement_mutation>(4,5)<complement_mutation><alter_mutation>(0,1)<alter_mutation><chromosome><chromosome><complement_mutation>(2,3)<complement_mutation><alter_mutation>(2,3)<alter_mutation><chromosome><mutation><round><round><reshape>4<reshape><crossover><type>single_point_crossover<type><single_point>3|0|<single_point><crossover><mutation><mutation_table>{'G':'A','A':'G','T':'C','C':'T'}<mutation_table><chromosome><complement_mutation>(3,5)<complement_mutation><alter_mutation>(2,3)<alter_mutation><chromosome><chromosome><complement_mutation>(4,5)<complement_mutation><alter_mutation>(3,3)<alter_mutation><chromosome><chromosome><complement_mutation>(5,6)<complement_mutation><alter_mutation>(0,0)<alter_mutation><chromosome><chromosome><complement_mutation>(4,5)<complement_mutation><alter_mutation>(1,2)<alter_mutation><chromosome><chromosome><complement_mutation>(3,5)<complement_mutation><alter_mutation>(2,2)<alter_mutation><chromosome><mutation><round><round><reshape>10<reshape><crossover><type>both<type><rotate><rotation_offset>6<rotation_offset><rotation_types>left|right|<rotation_types><rotate><single_point>8|<single_point><crossover><mutation><mutation_table>{'G':'A','A':'G','T':'C','C':'T'}<mutation_table><chromosome><complement_mutation>(13,15)<complement_mutation><alter_mutation>(2,4)<alter_mutation><chromosome><chromosome><complement_mutation>(14,16)<complement_mutation><alter_mutation>(0,4)<alter_mutation><chromosome><mutation><round>

B. Text Encryption Results

The proposed system can support Image and text encryption under a single framework. Text encryption begins upon selecting the "Text" radio button followed by clicking the Process Encryption and Decryption Operation button, as shown in Fig. 3.

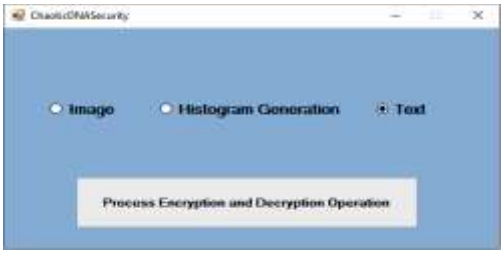


Fig. 3. User Interface to Select Text to Process Encryption and Decryption Operation.

In this proposed approach DNA cryptographic functions are designed and implemented with the use of Python script. The front end and the algorithm calling functions are designed using C # language written over Microsoft Visual C # Express. The following results were obtained on running python scripts over Spyder (Python 3.6) platform.

Case 1:

Encryption Process

Text: Hello

Key: 128 bits

00111000010010100101001101000110001100010101000001
01100101100101011010110011010001000001010010010011
00100100100101001111010011 00

Initial DNA sequence: CAGACGCCCGTACGTACGTT

Final DNA sequence: CTTAACACGTAGCTCCAGTA

Decryption Process

Encrypted text: CTTAACACGTAGCTCCAGTA

Key: 128 bits

00111000010010100101001101000110001100010101000001
01100101100101011010110011010001000001010010010011
00100100100101001111010011 00

Initial DNA sequence: CTTAACACGTAGCTCCAGTA

Decrypted text: Hello

Case 2:

Encryption Process

Text: Hello

Key: 128 bits

01010010001100110011010101101010001100110100100101
00111000110100011100100110111101010001010001110111
01010100011101000111001110 00

Initial DNA sequence: CAGACGCCCGTACGTACGTT

Final DNA sequence: TGAGTGGTTCCGCAAGCAGA

Decryption Process

Encrypted text: TGAGTGGTTCCGCAAGCAGA

Key: 128 bits

01010010001100110011010101101010001100110100100101
00111000110100011100100110111101010001010001110111
01010100011101000111001110 00

Initial DNA sequence: TGAGTGGTTCCGCAAGCAGA

Decrypted text: Hello

The above example illustrates the proposed model for text encryption using DNA sequences with a text sample as "Hello". First, the key will be computed then it can be used to produce initial and final DNA sequences. Meantime, DNA encoding rules along with single-point crossover, mutation, and complementary rules are used to obtain encrypted text [46]. During the decryption process, the same procedure is repeated and reversed, with the same key value as shown in Case 1 [47].

When there is another communication session with the same text input as "Hello" as shown in Case 2, then it computes a key value that is different than the previous session key. This will bring the proposed approach to high dynamicity and randomness. Since the keys generated and used were different during the various sessions/transactions makes it difficult to access the key computationally. Hence plaintext recovery is infeasible for an attacker. Table VII shows time required values in seconds for encryption and decryption processes. The time required for encryption and decryption is often found to be comparatively closer and takes less time. The decryption needs little more time than the encryption.

Table VIII shows the memory allocation (in bytes) of the proposed DNA Cryptographic method for encryption on disk. Table values indicate there is no need for more memory than the input file size.

TABLE VII. TIME REQUIRED VALUES

Case	Encryption Time(sec)	Decryption Time(sec)
Case 1	0.0032889842987060547	0.007483005523681641
Case 2	0.006028413772583008	0.009949922561645508

TABLE VIII. MEMORY REQUIREMENT FOR ENCRYPTION

Input Text size	Input file Size on disk	Encrypted file Size	Encrypted file Size on disk
86 bytes	4 KB (4096 bytes)	4.03 KB (4,128 bytes)	8.00 KB (8,192 bytes)
996 bytes	4.00 KB (4,096 bytes)	46.6 KB (47,808 bytes)	48.0 KB (49,152 bytes)

V. Conclusion

The proposed new DNA based cryptographic framework provides security to both image and text. The proposed approach can resist Differential attack, Brute Force attack, Chosen Plaintext attack, Dictionary attack, and other Statistical Attacks. The experimental results have shown that the histograms of encrypted images are uniformly distributed. Correlation coefficient values are found to be smaller in one or more directions. The proposed image encryption technique is better than existing in terms of information entropy, NPCR, and UACI values. Information entropy value for Peppers image is found to be 7.999182, which is 0.30% improvement over existing works. NPCR and UACI values for Lena image are 99.6279 and 49.7571, respectively. UACI value of the proposed method shows 1.06% improvement for encrypted Lena image over existing approaches. These parameters have shown that the proposed method is stronger. Proposed work support image files such as .jpg, .png, jpeg, .tiff formats, and text characters as input. In addition to Image encryption, there can be a text encryption option is also provided under the same framework. Compared with conventional algorithms, the overall execution time is considerably reduced in the proposed text encryption method. For the suggested solution, space consumption is less. In the future, this system may include audio encryption and can support very large files with different file formats for encryption and decryption processes.

Acknowledgment

I Bahubali Akiwate take this opportunity to express my deep sense of gratefulness to my guide and mentor Dr. Latha Parthiban, Pondicherry University, Pondicherry, India for her valuable advice, expert guidance, and unceasing support at all the stages during this research work. I extend my sincere gratitude to Dr. Veena Desai, Gogte Institute of Technology, Belagavi, Karnataka, India for her timely suggestions and encouragement in accomplishing this work.

References

[1] M.A. Ben Farah, R. Guesmi, A. Kachouri, M. Samet, "A novel chaos based optical image encryption using fractional Fourier transform and DNA Sequence operation" Optics and Laser Technology 121 (2020) 105777.

[2] Kang Xuejing, Guo Zihui, "A new color image encryption scheme based on DNA encoding and spatiotemporal chaotic system", Signal Processing: Image Communication 80 (2020) 115670.

[3] K.C. Jithin, Syam Sankar, "Colour image encryption algorithm combining Arnold map, DNA sequence operation, and a Mandelbrot set", Journal of Information Security and Applications 50 (2020) 102428.

[4] Je Sen Teh, Moatsum Alawida, You Cheng Sii, "Implementation and practical problems of chaos-based cryptography revisited" Journal of Information Security and Applications 50 (2020) 102421.

[5] Qiang Zhang, Ling Guo, Xiaopeng Wei, "Image encryption using DNA addition combining with chaotic maps", Mathematical and Computer Modelling 52 (2010) 2028-2035.

[6] Ivan Jiron, Susana Soto, "A new DNA-based model for finite field arithmetic" Heliyon 5 (2019) e02901.

[7] Ahmed M. Elshamy, Aziza I. Hussein, "Color Image Encryption Technique Based on Chaos", Procedia Computer Science 163 (2019) 49–53.

[8] Said HRAOUI, Faiq Gmira, "A New Cryptosystem of Color Image Using a Dynamic-Chaos Hill Cipher Algorithm", Procedia Computer Science 148 (2019) 399–408,

[9] Saswat K Pujari, Gargi Bhattacharjee, "A Hybridized Model for Image Encryption through Genetic Algorithm and DNA sequence", Procedia Computer Science 125 (2018) 165–171.

[10] N. Sasikaladevi, K. Geetha, A. Revathi, "EMOTE – Multilayered encryption system for protecting medical images based on binary curve", Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.01.014.

[11] Manreet Sohal, Sandeep Sharma," BDNA-A DNA inspired symmetric key cryptographic technique to secure cloud computing", Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.09.024.

[12] Bhaskar Mondal, Tarni Mandal, "A light weight secure image encryption scheme based on chaos & DNA computing", Journal of King Saud University – Computer and Information Sciences (2017) 29, 499–504.

[13] F.J. Farsana, V.R. Devi, "An audio encryption scheme based on Fast Walsh Hadamard Transform and mixed chaotic key streams", Applied Computing and Informatics, https://doi.org/10.1016/j.aci.2019.10.001.

[14] Mumthas S, Lijiya A, "Transform Domain Video Steganography Using RSA, Random DNA Encryption and Huffman Encoding", Procedia Computer Science 115 (2017) 660–666.

[15] Md. Rafiul Biswas, "A technique for DNA cryptography based on dynamic mechanisms", Journal of Information Security and Applications 48 (2019) 102363.

[16] Junxin Chen, "Exploiting self-adaptive permutation–diffusion and DNA random encoding for secure and efficient image encryption", Signal Processing 142 (2018) 340–353.

[17] Xingyuan Wang, Yu Wang, "A novel chaotic algorithm for image encryption utilizing one-time pad based on pixel level and DNA level", Optics and Lasers in Engineering 125 (2020) 105851.

[18] Wenjian Gao, Jie Sun, "Digital image encryption scheme based on generalized Mandelbrot- Julia set", Optik - International Journal for Light and Electron Optics 185 (2019) 917–929.

[19] Yu-Guang Yang, Bo-Wen Guan, "Image compression-encryption scheme based on fractional order hyperchaotic systems combined with 2D compressed sensing and DNA encoding", Optics and Laser Technology 119 (2019) 105661.

[20] Hossein Movafegh Ghadirli, "An overview of encryption algorithms in color images", Signal Processing 164 (2019) 163–185.

[21] Rasul Enayatifar, "Index-based permutation-diffusion in multiple-image encryption using DNA sequence", Optics and Lasers in Engineering 115 (2019) 131–140

[22] Shuliang Sun, "A Novel Hyperchaotic Image Encryption Scheme Based on DNA Encoding, Pixel-Level Scrambling and Bit-Level Scrambling" IEEE Photonics Journal, Volume 10, Number 2, April 2018.

[23] Taiyong Li, Minggao Yang, "A Novel Image Encryption Algorithm Based on a Fractional-Order Hyperchaotic System and DNA Computing", Hindawi Complexity Volume 2017, Article ID 9010251, https://doi.org/10.1155/2017/9010251.

[24] Z. Azimi S. Ahadpour, "Color image encryption based on DNA encoding and pair coupled chaotic maps" Multimedia Tools and Applications https://doi.org/10.1007/s11042-019-08375-6, part of Springer Nature 2019.

[25] Hatem M. Bahig, "DNA-Based AES with Silent Mutations", Arabian Journal for Science and Engineering (2019) 44:3389–3403.

[26] Oinam Bidyapati Chanu, " A survey paper on secret image sharing schemes", International Journal of Multimedia Information Retrieval (2019) 8:195–215. (part of Springer Nature 2018).

[27] Siyamol Chirakkarottu, Sheena Mathew, "A novel encryption method for medical images using 2D Zaslavski map and DNA cryptography", https://doi.org/10.1007/s42452-019-1685-8, Springer Nature Switzerland AG 2019.

[28] Ahmed Elhadad, "Data sharing using proxy re-encryption based on DNA computing", Soft Computing https://doi.org/10.1007/s00500-019-04041-z, Springer-Verlag GmbH Germany, part of Springer Nature 2019.

[29] M. A. Ben Farah, A. Farah, “An image encryption scheme based on a new hybrid chaotic map and optimized substitution box”, Springer Nature B.V. 2019.

[30] Manjit Kaur, Vijay Kumar, “A Comprehensive Review on Image Encryption Techniques”, Archives of Computational Methods in Engineering (Springer 2018).

[31] Xingyuan Wang, Huaihuai Sun, ” A chaotic image encryption algorithm based on zigzag-like transform and DNA-like coding”, Multimedia Tools and Applications (2019) 78:34981–34997, Part of Springer Nature 2019.

[32] Xiangjun Wu, ” Lossless chaotic color image cryptosystem based on DNA encryption and entropy” ,Nonlinear Dyn (2017) 90:855–875 DOI 10.1007/s11071-017-3698-4.

[33] Kareem Ahmed, Ibrahim El-Henawy, “Increasing robustness of data encryption standard by integrating DNA cryptography”, International Journal of Computers and Applications, VOL. 39, NO. 2, 91–105, 2017.

[34] P. K. Naskar & A. Chaudhuri, “Secured secret sharing technique based on chaotic map and DNA encoding with application on secret image”, The Imaging Science Journal, 64:8, 460-470, DOI: 10.1080/13682199.2016.1239427, 2016.

[35] Shahna k.U., Anuj Mohamed, “A novel image encryption scheme using both pixel level and bit level permutation with chaotic map”, Applied Soft Computing Journal 90 (2020) 106162.

[36] Junxin Chen,Lei Chen, Yicong hou, “Cryptanalysis of a DNA-based image encryption scheme”, Information Sciences 520 (2020) 130-141.

[37] Abdorreza Babei, Homayun Motameni, “A new permutation-diffusion-based image encryption technique using cellular automata and DNA sequence”, Optik-International Journal for Light and Electron Optics 203 (2020) 164000.

[38] Yujia Liu, “Optical image encryption algorithm based on hyper-chaos and public-key cryptography”, Optics and Laser Technology 127 (2020) 106171.

[39] Hongye Niu — Changjun Zhou, “Splicing Model And Hyper–Chaoti System For Image Encryption”, Journal of ELECTRICAL ENGINEERING, VOL 67 (2016), NO2, 78–86.

[40] Wei Feng, Yigang He. "Cryptanalysis and improvement of the hyper-chaotic image encryption scheme based on DNA encoding and scrambling", IEEE Photonics Journal, 2018.

[41] Shuliang Sun, Yongning Guo, Ruikun Wu. "A Novel Image Encryption Scheme Based on 7D Hyperchaotic System and Row-column Simultaneous Swapping", IEEE Access, 2019.

[42] K. Abhimanyu Kumar Patro, Bibhudendra Acharya, Vijay Nath. "Various dimensional colour image encryption based on non overlapping block-level diffusion operation", Microsystem Technologies, 2019.

[43] Grasha Jacob, Murugan Annamalai. "DNA Sequence Based Cryptographic Solution for Secure Image Transmission", IGI Global, 2016.

[44] Wang, Xing-Yuan, Ying-Qian Zhang, and Xue- Mei Bao. "A novel chaotic image encryption scheme using DNA sequence operations", Optics and Lasers in Engineering, 2015.

[45] Sohal M, Sharma S, "DNA Inspired Symmetric Key Cryptographic Technique to Secure Cloud Computing", Journal of King Saud University - Computer and Information Sciences, 2018.

[46] Ying-Qian Zhang, Yi He, Pi Li, Xing-Yuan Wang. "A new color image encryption scheme based on 2DNLCML system and genetic operations", Optics and Lasers in Engineering, 2020.

[47] Mohammad Seyedzadeh, S.. "A fast color image encryption algorithm based on coupled two dimensional piecewise chaotic map", Signal Processing, 201205.

[48] Roayat Ismail Abdelfatah. "Audio Encryption Scheme Using Self-Adaptive Bit Scrambling and Two Multi Chaotic-Based Dynamic DNA Computations", IEEE Access, 2020.

Smart Control System for Smart City using IoT

Parasa Avinash[1], B Krishna Vamsi[2], Thumu Srilakshmi[3], P V V Kishore[4]*
Department of Electronics and Communication Engineering Koneru Lakshmaiah Education Foundation, Guntur, INDIA

***Abstract*—As technologies are introducing and improving day by day, there is a tremendous change in the applications like "Smart City". The Internet of Things (IoT) is the best approach to combine various Sensors with Embedded devices to create solutions for the real time problems and this will help us to connect with Internet Society. The term IoT means controlling the things through Internet, in other terminology the objects will "talk" to each other and build some communication to work or to react. There are so many attention-grabbing modules in our society, so in this project we will Implement a model of Smart City with around six elements like Smart Garbage System, Smart Irrigation System, Smart Building, Smart Parking System, Restaurant Menu Ordering System and Manhole Detection and Monitoring System that too in an advanced way. Which means we are going to make some advancement in all these elements like Sending messages by using GSM module with Sensor responses, Sending the information and controlling the components through cloud platform like ADAFRUIT, Accessing webpages through IP Address using Networking domain, Enabling Technologies, Connectivity models and we are going to make all this system automatic. The main objective of this project is without any Human Involvement all the systems or elements has to work to make life easier. The required technologies in this project is Internet of Things (IoT), Embedded Systems and Networking.**

***Keywords*—Internet Society; Attention-Grabbing; Networking; Enabling Technologies; Connectivity Models; IP Address; ADAFRUIT; Embedded Systems**

I. Introduction

A. Smart Garbage Monitoring System

In this module we are going to implement a dustbin which will automatically opens the lid when some person reaches towards it, this could be done by using measuring the distance between the person and dustbin. Second aim is to create a webpage using ESP-01 module with reference to the default IP address "192.168.4.1" to store the status of the bin like how much area is filled. And if once the bin is full it will display in the webpage.

B. Smart Parking System

The main aim of this module aim is to store all the details related to parking place in the could platform called ADAFRUIT. And also to control the entry and exit gates though that platform, it will also maintain the slot timings like then the car is parked and when the car leaves the parking slot.

In this module, first we are going to create a project in ADAFRUIT cloud platform. That project will keep the information about the parking slot like Entry Gate, Exit Gate, Cars Parked, Entry Slot 1, Exit Slot 1, Entry Slot 2, Exit Slot 2, Entry Slot 3, Exit Slot 3. At each and every instance of time that page will update to display the information that was executing in the parking slot. So that we can able to look over the whole process that is going on in the parking slot without any physical appearance. And another thing is once the car appeared at the entrance we can able to open the entry gate automatically by clicking a button on that page created by us, and the same thing was implemented for exit gate also. And another element in this module is special parking, which means if once the car is parked in that slot the owner will get a message like "Your car is parked successfully" and if the car was left from that place the owner will get another message like "Slot is EMPTY".

C. Smart Building System

First we are going to implement smart window curtain and smart water pump now let us discuss about the smart water pump this project is aim to design and construct the light sensing blinds to achieve the aforementioned goals. Hardware test results from this project demonstrate the capabilities of smart blind system to measure ambient light inside the room and outside the window, to adjust the angle of each of the blinds on the window, and to change the desired brightness of the room. In the second submodule, based on the sensor values the tap will automatically open when it detect binary digit "1" in the values and the tap was closed when it detects the binary digit "0".

D. Manhole Detection System

In the existing work of the manhole detection system using IOT they are several drawbacks and many people fall in manhole and die but using this project we can stop that our idea is to create a webpage and sends the information to the person near the Manhole. We are doing this using the GSM module. GSM module sends the message using the webpage. Another important aspect in this module is creating a webpage to store the details related to manhole. This will help to have a complete analysis on the manhole data.

E. Smart Irrigation System using ADAFRUIT IO

The world is changing as time and so on agriculture. Nowadays, People are integrating electronics in every field and agriculture is not an exception to this. This merging of electronics in agriculture is helping farmers and people who manage gardens. In this article, we will see how to monitor and how to manage gardening and agriculture. We will use (ESP32) controlling module for IoT and we will update the data on the cloud and based on readings we will take the appropriate action. In this project we have used sensors like LDR(Light dependent Resistor), Temperature sensor, Soil Moisture level sensor and we will use the water pump to react on the sensor's data. Apart from this, we can use lots of sensors to monitor.

*Corresponding Author

F. Restaurant Menu Ordering System using Webserver

The main theme of the project is to display the particular item which is ordered by the user and that item should display on the web server with the IP address and hence the web page should consists of the item name , item cost and finally it should print the total cost of item on the webserver by using ESP8266 it can import the data or gather the data through the web server on the transmitter side if the user gives the order of an particular item it should display on the receiver side i.e., on 16*2 display and through web server. At last by using the ESP8266 we can transmit the data through web server and it mainly creates the default IP address on html page such as 192.168.2.1 on that it should display the menu. And for suppose if the user gives the order of a particular item it should display the item name and item cost and total cost of an item through webserver. In this project the user uses the TFT Touch display shield to select the items i.e., when he selected the item it will display the item and item cost on the 16*2 LCD display. Creating a webpage using a hardware module like ESP-01 is a new techniques which will help us to monitor/to see data on webpage, and this webpage will be created with reference to default IP address. The food ordering system is proposed with the use of a handheld device placed on each table which is used to make an order at the restaurant. The system uses a TFT touch plus LCD display module which is placed on each customers table for them to make orders. Order is made by selecting the items displayed on LCD.

II. Related Work and Advancements

A. Drawbacks in Existing Works of Smart Garbage Monitoring System (Fig. 1)

1) The main drawback is, in this element Human Involvement is needed for some tasks like to clear dust/over flow.

2) Major part of this element depends upon the working of the Wi-Fi module which is operable only in small distances.

3) All the existing works only focus on any one of the perspective like Household or the public places, they don't aim on both the scenarios at a time.

4) For example some of the projects only focusing on intimating the status of bin through mobile app or LCD display in household and another scenario is intimating through message to the municipal officer regarding the status whether full or not in some areas.

5) It is not possible to display the quantity filled in the bin through long distances.

Advancements in this module:

a) The advancement which we are implementing in this element is like sending messages by using GSM module sensor responses.

b) And another important domain used in the project is "NETWORKING". We are going to access this element by using IP address.

c) Which means that we are going to create a webpage and assigning some setup with IP address so that we can able to control/access all these elements at anytime and from anywhere by storing them in a database.

d) The creating of webpage is done by using hardware element like ESP-01, which is modern technique to create a webpage with reference to IP address.

B. Drawbacks in Existing Works of Smart Parking System (Fig. 2)

1) In the existing works there is no mechanism like sending the SMS to mobile once the car is parked and again the car is moved from the parking place.

2) In existing projects, no one use the specific application to store the information related to parking like ADAFRUIT, Webpages, etc.

3) The time slot is not mentioned in existing works like, when the car is placed and when it is moved from the parking.

Advancements in this module:

a) When the car is placed in the parking place, then one SMS is sent to our mobile like car is placed at position 4. So that no need to go and check where the car is placed.

b) When the car left from the parking place also we will get a message to our mobile, this will provide a security to our car.

c) In this project I am using an software/application called ADAFRUIT, in that I have created a project mentioning like Number of cars parked, Entry gate(open/close), Exit gate(Open/close), Slot1, Slot2 and Slot3. They will mention all the details of the parking place.

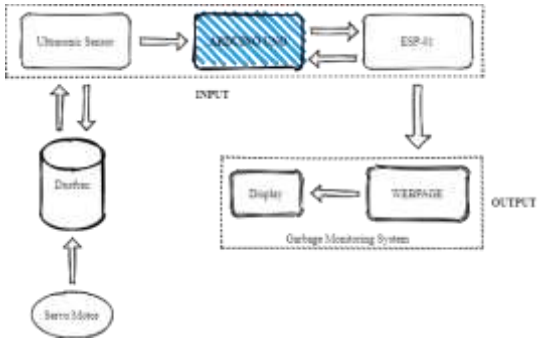


Fig. 1. Block Diagram for Smart Garbage Monitoring System.

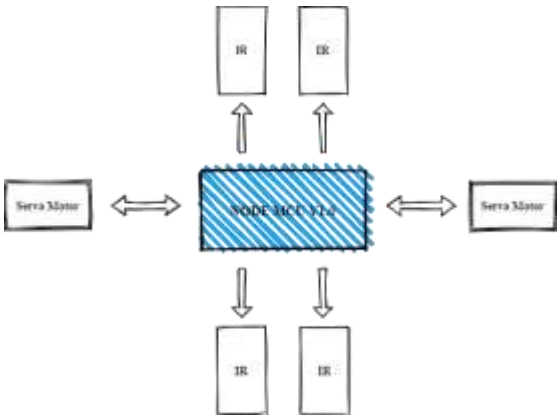


Fig. 2. Block Diagram for Smart Parking System.

C. Drawbacks in Existing Works of Smart Building System (Fig. 3)

1) The disadvantages of a new technology are just as important as the advantages, if not more important. People who are not computer-savvy would not want to live in a smart home since it is completely computer run. People who are not computer-savvy would not want to live in a smart home since it's completely computer run. If someone were to hack a home's system they would have access to controlling the home, like unblocking doors, monitoring the cameras off so they would invalid the home.

2) In existing works I observed that the watering of plants through sprinklers, this mechanism is not implemented for handwash which will reduce the human effort.

Advancements in this module:

a) This project is solving the problem of wasted energy within buildings and homes, because currently the lights turned on inside building do non-utilized natural, ambient light from the sun. Rather than having unnecessary light from a light source, the automated light sensing smart blinds can sense the amount of light outside the window and in the room, and then adjust the angle of the blinds to save energy by utilizing the available outdoor light. This way, the light source will not be running at maximum power output while there is excess light coming through the window.

b) Coming to the tap mechanism, here the human involvement is not needed to open/close the tap. In previous works I observed that watering the plants using sprinklers, based on that mechanism I design this module of automatic open/close the tap.

D. Drawbacks in Existing Works of Manhole Detection System (Fig. 4)

1) Manholes leading to underground supply systems are essential for their maintenance, for example, it concerns telecommunication networks, water supply networks, gas supply networks and electricity networks, and so on. Although it is very crucial to a city's operations, the manhole can be one of the least protected and most vulnerable assets.

2) In previous works, the message sending mechanism is not included which is the main drawback. Now a days mobile is mandatory to every person so, by intimating through an SMS is needed.

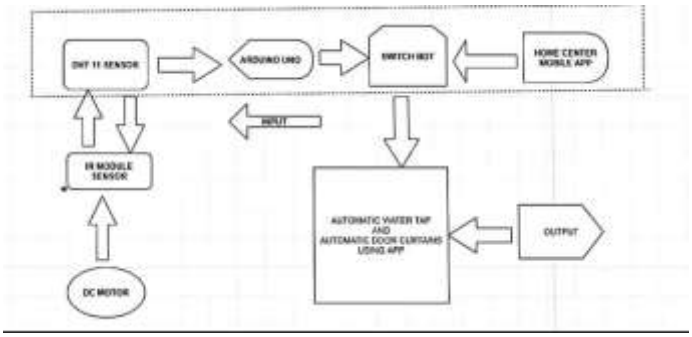


Fig. 3. Block Diagram for Smart Building System.

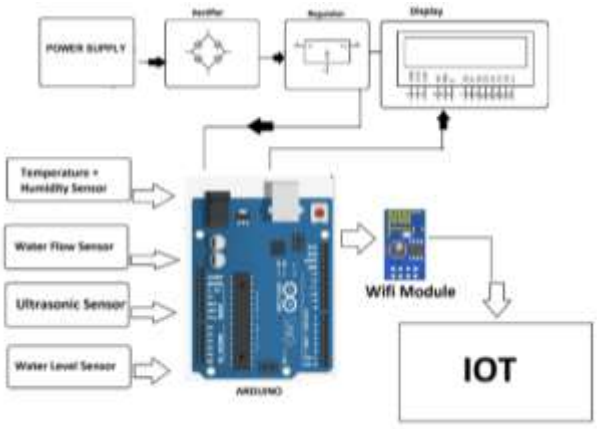


Fig. 4. Block Diagram for Manhole Detection System.

Advancements in this module:

a) The advancement which we are implementing in this element is like sending messages by using GSM module sensor responses.

b) We control and detect man hole by using GSM module sensor. GSM module sensor detect the person near the manhole and sends information to him.

c) We create a webpage to store the information of the person.so when the person comes near the man hole GSM module takes the information and message to him/her by using a webpage. We took up this project to control the falling of people in man hole.

E. Drawbacks in existing works of Smart Irrigation System Using ADAFRUIT (Fig. 5)

1) The main drawback of this project is without any human involvement all systems has to work make life easier to farmers.

2) Actually the existing work in smart irrigation system is with the human involvement we can control the plants when it is wet or dry state conditions.

3) The another type of existing work in smart irrigation system is using BLYNK app in this with only human involvement it can control the plants whether the plant is in dry or wet condition state.

4) The brilliant farming necessities accessibility of web ceaselessly. Provincial piece of the greater part of the non-industrial nations don't satisfy this prerequisite. Also web association is much slower.

Advancements in this module:

a) In this module we will see the monitoring and how to manage gardening of agriculture. And in this module we will use the ESP32 module for controlling the garden.

b) In this part we will update the data on the cloud and based on the readings we will take the accurate readings from it.

c) In this project we used LDR, temperature sensor, soil moisture sensor and submersible water pump to react on sensor data.

d) If the soil moisture level is very low then it will turn ON the Water Pump. We are monitoring the motor status as well for the feedback to confirm the motor status.

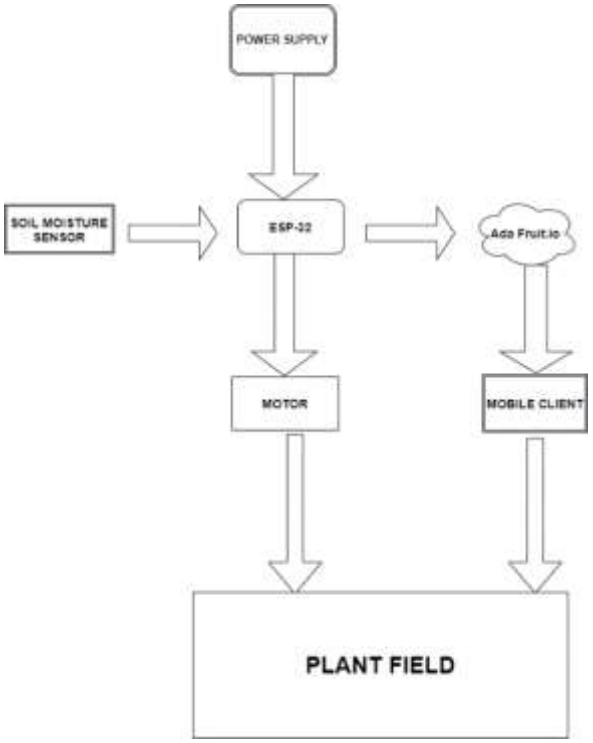


Fig. 5. Block Diagram for Smart Irrigation System.

F. Drawbacks in Existing Works of Restaurant Menu Ordering System (Fig. 6)

1) In existing Work the user uses the TFT Touch display shield to select the items i.e. when he selected the item it will display the item and item cost on the 16*2 LCD display.

2) The second aim is to create a web page using Esp8266 module with that reference to the default Ip address "192.168.2.1" is used to store the status of the items and cost of the items and total cost of an particular item once if we select a particular item it should display the total cost of the food item and name of the food item through particular web page.

Advancements in this module:

a) Firstly when the user was selected particular item the system should display the selected item name, cost of the item, and at last it should display the total cost of an selected items through the IP address (web server).

b) In this project I am using ESP8266 to import the data or gather the data through the webserver hence with ESP8266 module with that module reference it creates the default Ip address "192.168.2.1" which is used to store the "name of the item", "cost of item", and total cost of all items which we have selected it should display on web page.

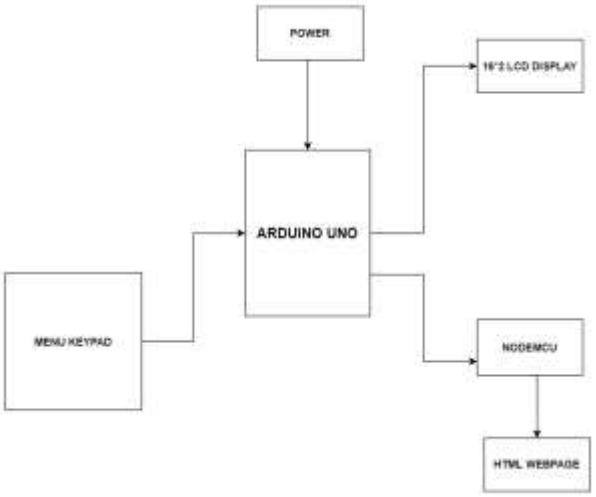


Fig. 6. Block Diagram for Restaurant Menu Ordering System.

III. Results and Discussions

A. Smart Garbage Monitoring System

The first output for this module is lid opening by measuring the distance of the object which is placed infront of it, after opening the lid it will take some delay and automatically closed. This delay time will help us to through the dust in the bin. Secondly, we have to connect to the WIFI module which will appear in out device so that it will create a webpage by using default IP address in the device which was connected to WIFI module. In that webpage we can able to monitor the level of the dustbin and once it was full it will display a message like "Dustbin is full" (see Fig. 7 to 9).

Fig. 7. Opening of Lid.

Fig. 8. Connecting to WIFI.

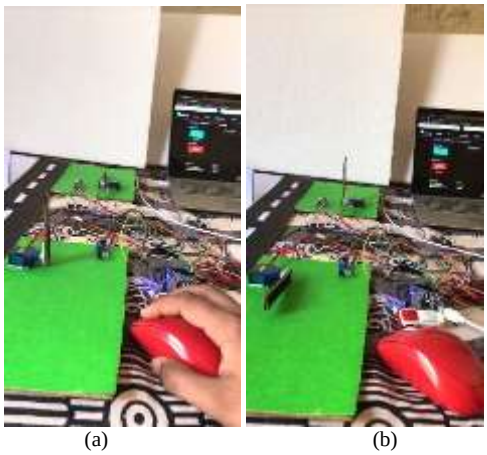
Fig. 9. Webpage Display for Smart Garbage Monitoring System.

B. Smart Parking System

The first output for this module is to open the entry and exit gates remotely by using the adafruit platform this will help to validate the customer whether he paid the parking fee or not. Secondly, the entry and exit gates has to open automatically when the car is placed infront of the gates. Third and most important output is to display the slot timing of the cars which was placed in parking slots. The number of cars count which enters/leaves the parking place will be updated in the cloud service platform called ADAFRUIT. Another interesting submodule is special parking, when the car was parked in this slot the owner will get a message to their mobile like "Your car is parked successfully !!" and once the car was removed from that slot they will get another message like "Slot is empty !!" this will provide more safety to the car. And based on all these measurements that ADAFRUIT platform will form a graph so that we can have more understanding about each and every slot (see Fig. 10 to 13).

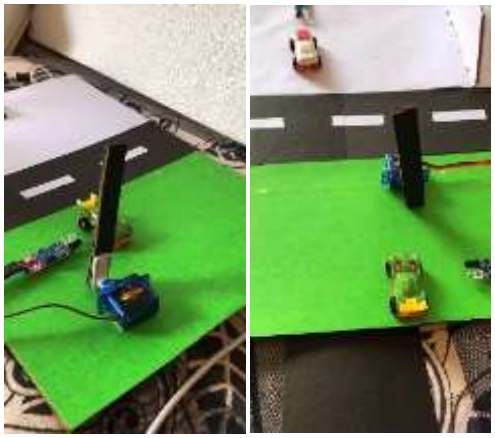
C. Smart Building System

The first output of the module is automatic open/close of a curtain. When we clap the window curtain will open and when we clap for another time windows will close (see Fig. 14). And second one is smart water pump when we place hand under the tap, it will automatically on and once we remove the hand under the tap, it will off automatically (Fig. 15). In automatic curtain we use a webpage to record the time how much time did the curtain is opened (Fig. 16). And in smart water pump we record the amount of water by using ESP11 WIFI module and we record it using a webpage.



(a) (b)

Fig. 10. Entry and Exit Doors Open through ADAFRUIT.



(a) (b)

Fig. 11. Automatic Entry and Exit Doors Opening.

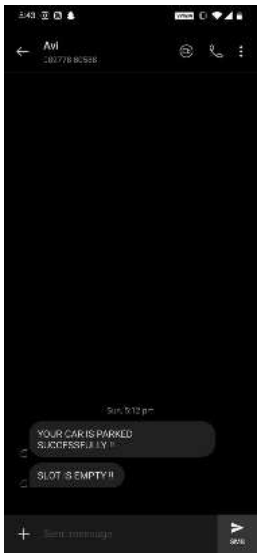


Fig. 12. SMS to Mobile for Special Parking.

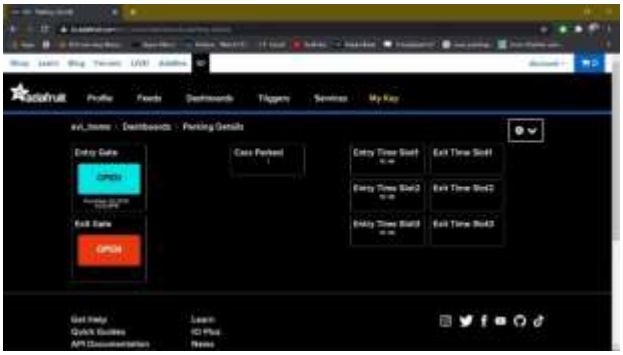


Fig. 13. Adafruit Display for Smart Parking Display.

(a) (b)

Fig. 14. Opened/Closed Window Curtain Automatically.

Fig. 15. Smart Water Pump.

Fig. 16. Automatic Water Flow in Smart Building System.

D. Manhole Detection System

The output of this module is to detect the manhole. Using this process we can reduce the people falling in manhole. In this project we used GSM module and a webpage the GSM module will detect the person near the manhole using ultra sonic sensor and the person will get the message to his mobile like “manhole detected” otherwise in the webpage it will be as “Rootcleard”. At first the output was displayed in LCD display they we can able to see the output in webpage also (Fig. 17). By alert message sent by the GSM will help us to avoid the accident cause by the manhole (Fig. 18).

E. Smart Irrigation System using Adafruit IO

The first output for this module is to water the plants through adafruit io setup firstly in this module we have to create the interfaces in adafruit such as temperature, light and soil moisture percentage (graph interface) and atlast we have to create an interface called motor which is used to on/off if the plant condtion is in dry then the motor should on in the setup and if the plant condtion is wet then the motor should off and hence it should show the soil moisture percentage through graph type and it should show the light condition around the plant field. Hence the theme of this project is without any human involvement it should help the farmers to water the plant field (see Fig. 19 and 20).

F. Restaurant Menu Ordering System

The main theme of this module is to print the item names and item cost and total cost of the items should print on the webserver i.e. (default IP address) firstly in this module ESP8266 is used to publish data through webserver and it is useful to print items which is given by the user and in this module Arduino uno was used and it is used to print the item names in 16x2 lcd display coming to the circuit, they are 5 push buttons which is used to display the items in 16x2 lcd as well as server IP address at last 5th pushbutton is used to print the total bill of the order which is given by the user (see Fig. 21 and 22).

Fig. 17. LCD Display for Manhole Detection.

Fig. 18. Detecting the Manhole.

Fig. 19. Automatic Water Flow in Smart Irrigation System.

Fig. 20. ADAFRUIT Display for Smart Irrigation System.

Fig. 21. LCD Display Coming to the Circuit.

Fig. 22. Webpage Display for Restaurant Menu Ordering System.

IV. Conclusion

The main objective of this project is without any Human Involvement all the systems or elements has to work to make life easier. The required technologies in this project is Internet of Things (IoT), Embedded Systems and Networking. IoT is the most emerging Technology with so many applications like Smart Home Automation, Health Care System via Mobile, Smart Grid, Smart Farming, Smart Surveillance, etc. So, the main concept is to create a Smart City with all the advanced elements which will reduce the Human efforts and trying to Implement this one in Real Life so that there is a lot of change in society.

V. Future Scope

All these modules are designed based on the real time problems faced in the society. These can further modified to an advanced levels as following:

Smart Garbage Monitoring System: This module can be further updated as garbage separation by using the AIML Techniques merge with the Internet of Things technology. In this the dry and wet waste can be separated automatically by a machine without any human involvement this could be done though camera detection.

Smart Parking System: This module can be further updated as a ticket counter, which means in my module the human involvement is needed to distribute tokens/tickets but we can implement a ticket counter which will automatically distribute it and for this high level of surveillance is also required.

Smart Building System: This module can be further implemented as automatic motor on/off system, which means the motor which will supply the water to tank can be automatically work based on the timings given to it or based on the sensor responses.

Manhole Detection System: This module can be further implemented as tracking the location of the manholes in different areas and storing the data in the databases, so that we can able to analyze the percentages of problems placed by the people due to manholes in particular area.

Smart Irrigation System: This module can be further implemented as, storing the data in the database so that we can able to analyze the position/condition of our garden/field in every season.

Restaurant Menu Ordering System: This module can be further extended as, displaying the table number in PC whenever a person press the button on a particular table. So that we can able to track the position of the table.

References

[1] Prof. Dr. Sandeep M. Chaware, Shriram Dighe, Akshay Joshi, Namrata Bajare, Rohini Korke, “Smart Garbage Monitoring System using Internet of Things (IOT)”, IJIREEICE, vol 5, Issue1, January 2017.

[2] Husam Rajab, Tibor Cinkelr, “IoT based Smart Cities”, The Institute of Electrical and Electronics Engineers, 2018.

[3] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, , and Michele Zorzi, “Internet of Things for Smart Cities”, IEEE Internet of Things Journal, Vol. 1, No. 1, February 2014.

[4] Miss. Priya A. Jadhao, Miss. Sonal D. Sakhare, Miss. Kajal G. Bhaldane, Prof. Abhishek P. Narkhede, Prof. Vaibhav S. Girnale, “Smart Garbage Monitoring and Collection System using Internet of Things”, A National Conference On Spectrum Of Opportunities In Science & Engineering Technology Volume 5, Special Issue 06, April-2018.

[5] Faheem, S.A. Mahmud, G.M. Khan, M. Rahman, H. Zafar, “A Survey of Intelligent Car Parking System”, Vol 11, Issue 5, 2013.

[6] Anusha, Arshitha M S, Anushri, Geetanjali Bishtannavar, “Review Paper on Smart Parking System”, IJERT, Special Issue - 2019.

[7] Husam Rajab,Tibor Cinkelr,"IoT based Smart Cites", 19-21 June 2018.

[8] Nor Adni Mat Leh, Muhammad Syazwan Ariffuddin Mohd Kamaldin ,Zuraida Muhammad,"Smart Irrigation System Using Internet of Things, 7 October 2019.

[9] Hamza BENYEZZA, Mounir BOUHEDDA ,Khaoula DJELLOUT,"Smart Irrigation System Based Thingspeak and Arduino 24-25 November 2018.

[10] Kriti Taneja , Sanmeet Bhatia ,"Automatic Irrigation System using Arduino UNO,15-16 June 2017.

[11] J.Caroline EL Fiorenza , Anurag Chakraborty, R Rishi, Kaustubh Baghel ,"Smart Menu Card System",15-16 Oct 2018.

[12] Renjith V Ravi,Amrutha N R,Amritha E,Haneena.P,Jaseena.T,"An Android Based Restaurant Automation System with Touch Screen",10-11 Jan 2019.

[13] Raviprakash Shriwas,Nikesh Patel, Asif Bherani,, Arti Khajone,Manish Raut,"Touchscreen Based Ordering System For Restaurants ,3-5 April 2014.

[14] XU Hongzhen,TANG Bin,SONG Wenlin,"Wireless Food Ordering System Based on Web Services ",10-11 Oct 2009.

[15] Wei, Z.; Yang, M.; Wang, L.; Ma, H.; Chen, X.; Zhong, R. Customized Mobile LiDAR System for Manhole Cover Detection and Identification. Sensors 2019.

[16] G.Gowtham, K.Hari Haran, G.Keerthee Rajan, A.Sweeto Jeison, “Sewage level maintenance using IoT” International Journal of Mechanical Engineering and Technical, vol. 9, Issue 2, February 2018.

[17] Dhanalakshmi.G, Akhil.S, Francisca Little Flower.M, Haribalambika.R, “Explosion detection and drainage monitoring system by Automation System” International Journal of Innovative research in computer and communication engineering, vol. 6, issue 2, February 2018.

[18] D.-Y., Ferranti, E. and Hadeli, H. (2013), “An intelligent building that listens to your needs”, Symposium on Applied Computing, Coimbra, March 18-23.

Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak

Mohammad Abu Kausar[1], Arockiasamy Soosaimanickam[2], Mohammad Nasar[3]
Department of Information Systems, University of Nizwa, Sultanate of Oman[1, 2]
Department of Computing & Informatics, Mazoon College, Sultanate of Oman[3]

***Abstract*—The COVID-19 pandemic, is also known as the coronavirus pandemic, is an ongoing serious global problem all over the world. The outbreak first came to light in December 2019 in Wuhan, China. This was declared pandemic by the World Health Organization on 11th March 2020. COVID-19 virus infected on people and killed hundreds of thousands of people in the United States, Brazil, Russia, India and several other countries. Since this pandemic continues to affect millions of lives, and a number of countries have resorted to either partial or full lockdown. People took social media platforms to share their emotions, and opinions during this lockdown to find a way to relax and calm down. In this research work, sentiment analysis on the tweets of people from top ten infected countries has been conducted. The experiments have been conducted on the collected data related to the tweets of people from top ten infected countries with the addition of one more country chosen from Gulf region, i.e. Sultanate of Oman. A dataset of more than 50,000 tweets with hashtags like #covid-19, #COVID19, #CORONAVIRUS, #CORONA, #StayHomeStaySafe, #Stay Home, #Covid_19, #CovidPandemic, #covid19, #Corona Virus, #Lockdown, #Qurantine, #qurantine, #Coronavirus Outbreak, #COVID etc. posted between June 21, 2020 till July 20, 2020 was considered in this research. Based on the tweets posted in English a sentiment analysis was performed. This research was conducted to understand how people from different infected countries cope with the situation. The tweets were collected, pre-processed and then text mining algorithms used and finally sentiment analysis have been done and presented with the results. The purpose of this research paper to know about the sentiments of people from COVID-19 infected countries.**

Keywords*—*COVID-19; corona virus; corona; pandemic; social media; sentiment analysis; Twitter

I. INTRODUCTION

Coronavirus disease (COVID-19) first was identified in December 2019 in Wuhan, China and has spread throughout the world covering every region. In 3-4 months this epidemic has disturbed the whole world. The world has witnessed many pandemic periods, but this pandemic today arouses severe economic problems on a country-scale as well as at micro scale. Individuals may experience psychotic symptoms due to pandemic and nations may suffer economic recession due to people with traveling restriction, and all activities pertaining to economics have been closed and social distancing has been imposed. Twenty one million people around the world were reported positive for COVID-19 by mid-August 2020 and nearly 773,072 were dead [1]. Now COVID-19 is increasing at very fast rate, especially in countries like USA and India. COVID-19 has affected more than 215 countries till 18th August 2020. The top 10 countries which have been severely affected by COVID-19 as on date 17th August 2020, includes USA (5,566,632 patients), Brazil (3,340,197 patients), India (2,647,663 patients), Russia (922,853 patients), South Africa (587,345 patients), Peru (535,946 patients), Mexico (522,162 patients), Colombia (468,332 patients), Chile (385,946 patients) and Spain (358,843 patients) [1].

Diverse use of social networking sites, like Twitter, speeds up the process of sharing information and having views on community events and health crises [2-5]. COVID-19 has been one of Twitter's trending areas throughout January 2020 and it has continued to be debated so far. Since more countries have adopted quarantine measures, people have increasingly relied on various social media sites to get news and expressing their opinion. Twitter data is useful in exposing public debates and feelings about exciting issues and real knowledge of emerging pandemics. In the ongoing COVID-19 pandemic, several government agencies around the world use Twitter as one of the key means of contact to frequently exchange policy updates and news related to COVID-19 with the general public [6]. Increasing numbers of studies have been collected from Twitter data since the COVID-19 outbreak to understand the general public's reactions and conversations related to COVID-19 [7-12]. For example, Abd-Alrazaq and colleagues use Tweets collected between 2nd February and 15th March 2020 to follow topic modeling and sentiment analysis to understand key topics and feelings around COVID-19 [7]. Doctors and individuals mentally more influenced by the epidemics are the most likely to speak about it on social networks like Twitter, which have become significant in our day to day lives.

The Twitter messages created via Twitter are named as Tweets. These data are available in public domain. It can thus be taken as raw data primarily for the extraction of opinions, for the analysis of customer fulfillment and for different rating policy schemes and, ultimately, a study of sentiment has been conducted. Even the online purchases nowadays take place on the basis of people's opinions about different products. For its positivity, advertisers and buying teams need to spend more time evaluating the consumer experience.

This research study has been conducted to identify the sentiments of the people of eleven different infected countries with COVID19 and identifies what emotions people have been sharing from different parts of the world. The countries selected for the study are top 10 [1] infected countries plus one more country from Gulf region, i.e. Oman. The countries selected for the study are USA, Brazil, India, Russia, South Africa, Peru, Mexico, Chile, Spain, UK and Oman.

II. Problem Statement

This research aims to capture, process and evaluate people's feelings within the certain timeframe on the tweets posted on twitter. The study would therefore concentrate on the following questions: i. Collect the tweets through Twitter API using RTweet package in R programming was used. The Hashtag used for collecting the tweet were #covid-19, #COVID19, #CORONAVIRUS, #CORONA, #StayHomeStaySafe, #Stay Home, #StayHomeSaveLives , #Covid_19, #CovidPandemic, #covid19, #CoronaVirus, #Lockdown, #Qurantine, #qurantine, #CoronavirusOutbreak and #COVID. ii. Preprocess the tweets by data cleaning (removing white spaces, links, punctuations, stop words, tokenization, retweet). iii. Calculate the sentiment using syuzhet package and analyze the result.

The tweets posted in English have been considered for a sentiment analysis to understand how people from different infected countries have responded during this pandemic situation to cope with it. The collected tweets will be used, pre-processed and applied with text mining algorithms for performing the sentiment analysis.

III. Related Work

Several researchers have been working on sentiment analysis on different social media data particularly on Twitter, few main contributions that help to discover user attitudes or sentiments in various cases when pandemic happening around the world. This section covers a number of the important papers, which was used as reference.

Researchers analyzed Twitter data for real-time projections of influenza spread and other communicable outbreaks [31]. Researchers measured the emerging risk in an outbreak of influenza in 2009 by analyzing tweet keywords and measuring the incidence of disease in real time and the efforts to prevent disease [32]. Throughout the 2014 outbreak of the Ebola virus, Twitter users shared important health information from media outlets with peak Twitter activities within 24 hours of the news events [33]. [13] Investigated the feelings concerning coronavirus COVID-19, thereby examining the feelings of various people about the pandemic. For this reason, the twitter API used to obtain useful corona virus tweets, and then analyzed based on positive, negative, and neutral emotions with the help of machine learning techniques. Additionally, authors used NLTK library for pre-processing of fetched tweets and the Textblob dataset has been used to evaluate the tweets, after that the exciting results indicates positive, negative, neutral feelings throughout various visualizations.

In [14], the researcher examined and visualized the effect of COVID-19 on the World by implementing certain machine learning methods and algorithms in sentiment analysis on the twitter dataset to recognize positive and negative views people across the globe. It shows that there has been stronger justification for the implementation of Naive Bayes' machine learning approach. In [15], the author drew up a list of COVID-19-related hash tags to looking for specific tweets during 14 days period from 14 to 28 January 2020. Tweets are collected via the API and stored in plain text form. Keywords associated with the level are identified and evaluated for instance, strategies for infection control; vaccination and racial discrimination were also analyzed. At last, the analysis on sentiment data to determine the emotional valence and predominant emotion of each tweet. Ultimately, over time, tweets are analyzed to identify with related topics using an unsupervised method of machine learning. [16] Illustrate observations into the development of anxiety-feeling over time as COVID-19 hit the highest levels in the United States, using textual descriptive analytics assisted by appropriate textual visualization. The author provides a conceptual insight of two important classification methods for machine learning throughout the field of sentiment analysis insights and compares their effectiveness in the classification of varying lengths of Coronavirus Tweets. Authors observe 91 percent classification accuracy for short Tweets using the Naïve Bayes process.

In [17], the authors proposed an effective platform to gather, store, manage, mine, and other activities, called MISNIS (Intelligent Mining of the Influence of Public Social Networks in Society). This program helps non-technical users to quickly mine data and has one of the highest levels of success in Portuguese language tweet collection. [18] Studied the emotional changes using Twitter posts. The understanding of the feeling associated with the text being evaluated [19] is some of the primary insights that can be gained from textual analytics. Twitter social media site includes knowledge which is rich and significant and used as a forum for expressing emotion among its users. Due to the immense number of views, [20] described about twitter is a multi-domain, which covers a wide variety of topics including: education, politics, which goods. One way of analyzing Twitter's large number of views is to apply sentiment analysis. Analysis of sentiment is an application of natural language processing, computer linguistics and text interpretation which classifies text into a division. [21] Sentimental analysis has several applications, for example in businesses, for reviews on products that allow businesses to understand feedback from users and social media reviews to analyze customer reviews. Opinion and sentimental mining were well investigated in this regard, and all the alternative techniques and research areas were discussed. In [22], author addresses Tree kernel and feature-based models used in twitter for sentimental analysis. [23] Reveals the seven (7) years of sentimental review of twitter. Since tweets on Twitter are a particular text not like a regular text, several other works tackle this concern, such as the work on short and concise texts. In [24], author evaluated the data with a large quantity of tweets that were taken as big data and therefore listed the words, sentences or whole records. Authors used the linear method to estimate tweet divisions. This analytical approach did better result and the accuracy was 85.23%. The tree-structured multi-linear principal component analysis (TMPCA) [41] proposed for text classification is a novel data processing technique. To facilitate the machine learning task that follows, the TMPCA can effectively decrease the size of the entire sentence data. In [42], the author proposed a new multi-modal attention (one for text and one for image) Unsupervised neural machine translation model that is trained under an auto-encoding and cycleconsistency paradigm. Tree-structured Multi-linear PCA (TM-PCA) reduces the size of input sequences and sentences, rendering the classification of sentences simple and fast. For text data classification, TM-

PCA with SVM has demonstrated better performance than recurrent neural network [43]. Neural networks may use information from all input sequences to predict each particular output element that is suitable [44].

IV. Research Method

Sentiment analysis in the micro-blogging domain is a relatively recent research field and a fair amount of relevant prior work on user reviews, papers, web posts, articles and general phrase level sentiment analysis has been done. These variations from twitter mostly due to the limit of 280 characters per tweet, which requires the user to express compressed opinion in very short text. Twitter is a platform for microblogging and social networking launched in March 2006. With 330 million active monthly users Twitter is the most popular and reliable social networking platform. Twitter has encouraged the researchers to determine the sentiments on almost everything, including sentiments towards public health information [25], digital technology [29], products [26], natural calamities [30], politics [28], movies [27] etc.

Between 21st June 2020 and 20th July 2020 we created a list of hashtags associated to COVID-19 to check for appropriate tweets. We retrieved the tweets using the advanced programming interface (API) of Twitter's search application and stored them as a CSV format. We carried out a sentiment analysis using tweet text to classify the emotional valence (positive, negative or neutral) of each tweet[34] and prevailing emotions (anger, disgust, fear, happiness, sadness, or surprise)[35]. Eventually, we did topic modeling using an unsupervised method of machine learning to classify and evaluate relevant topics over time within the tweet corpus [36].

A. Data Collection

From 21st June to 20th July 2020, about 1,305,000 tweets were collected from each infected country as shown in Fig. 1. For the tweet collection, RTweet package in R programming was used. The Hashtag used for collecting the tweet were #covid-19, #COVID19, #CORONAVIRUS, #CORONA, #StayHomeStaySafe, #StayHome, #StayHomeSaveLives , #Covid_19, #CovidPandemic, #covid19, #CoronaVirus, #Lockdown, #Qurantine, #qurantine, #CoronavirusOutbreak and #COVID; and the collected Tweets saved in CSV file. The retweets and replies were filtered out while collecting the tweets to avoid duplication of the tweets. As the complete database was obtained, the data cleaning process has been performed, where the white spaces, punctuation, stop words were removed. After the data cleaning process, the NRC Emtoion lexicon was applied with the help of get_nrc_sentiment function to analyze the tweets.

B. Sentiment Analysis

It is about measuring people's feelings, i.e. thoughts about a specific context such as product reviews etc. Sentiment analysis is the process whereby a portion of letters is positive, negative or neutral. A sentiment analysis system for text analysis incorporates natural language processing (NLP) and machine learning techniques to assign weighted feeling scores within a sentence or phrase to entities, topics, themes and categories. It is believed that the automatic sentiment analysis must also implement finely tuned algorithms to detail the human emotions. Mohammad and Turney [37] not just recognized the positive and negative lexical things, they likewise examined the basic feelings which has been characterized by Plutchik's eight fundamental feelings model. The NRC Word-Emotion Association Lexicon contains 10,170 lexical things which break down the positive and negative extremity as well as recognize the eight feelings characterized by Plutchik [38].

The 10,170 lexical items of the NRC include 1,587 most frequently used nouns, verbs, adverbs and adjectives, 640 words defined by Ekman subset from WordNet Affect Lexicon and 8,132 terms from General Inquirer. Syuzhet package version 1.0.4 [39] has implemented the NRC via open access with the method “src” and is freely available for language R. The syuzhet package has been improved over the years, following several issues raised by the researchers [40]. However, it was also stated that irony and sarcasm are two very complex emotions and are conveyed more on the basis of spoken texts rather than texts such as speeches.

The Syuzhet package in R has been used to compare four sentiment analysis algorithms: syuzhet (default), bing, afinn, and nrc. First, transform the clean text of a tweet into vectors to analyze tweets. A vector is a fundamental data structure in R that contains elements of the same kind, where words and phrases are the elements. When R recognizes a tweet as a variable, the sentiment analysis algorithms will independently evaluate and rate each word and expression. Here the vector function has been used to transform tweets into a vectors and forms the set of words and phrases into a new frame of data.

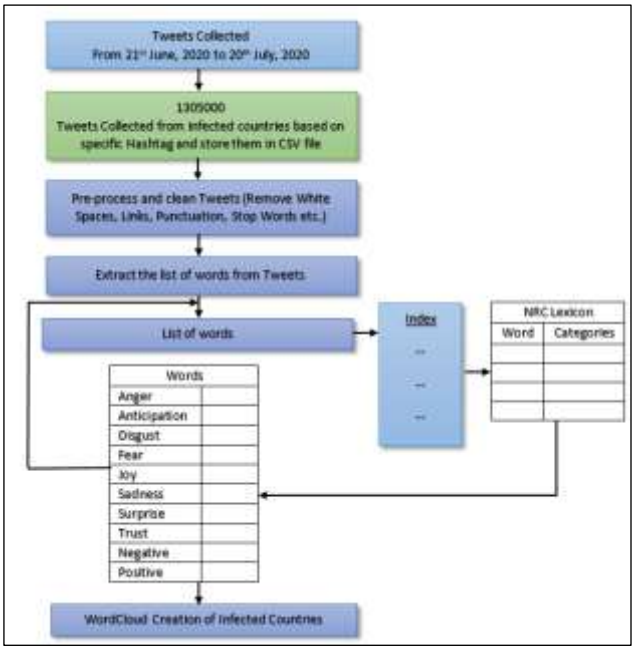


Fig. 1. Flowchart for Sentiment Analysis of the Tweets.

At the Nebraska Literary Lab, the Syuzhet lexicon was created and the name "Syuzhet" comes from the Russian formalists Victor Shklovsky and Vladimir Propp, who split the narrative into two parts, the "fabula" and the "syuzhet". The Syuzhet algorithm is used to evaluate literary works and focuses on how the text elements are constructed and assigns a sentiment score using fractions for each word ranging from -1 to 1. We used the get-sentiment function on the words-df data frame to extract the sentiment values on each tweet and bring the values into a new variable. On the basis of feelings (positive and negative), the syuzhet package classifies the tweets and categorizes them into 8 emotions (fear, joy, anticipation, anger, disgust, sadness, surprise, trust).

V. Experimantal Setup

R programming was used to collect the tweets through Twitter API using RTweet package. Many packages have functions for text classification and also for sentiment analysis in R programming. TM, tidytext, wordcloud, dplyr, syuzhet are some of the packages used. Applications for text mining use tm package in R. Tidytext is used to modify unstructured text data in such a way that it can be analyzed. The Wordcloud package has features that are used to build nice word clouds. Dplyr is a data manipulation grammar that offers a consistent collection of verbs to help you overcome the most common problems in data manipulation. From the syuzhet package, sentiment dictionaries, sentiment derived plots and feelings can be extracted. Datasets from the local library are imported. The collected dataset will be assigned to a corpus variable that can be used in R for preprocessing.

Prior to further processing with the text, all text must be preprocessed. To delete unused document type entry datasets, some text preprocessing methods were used. In text preprocessing, there are several techniques available, only some of them model have been used in this work. Special characters like @, #, / ... have no value adding to the sentiments of the review. The Term Document Matrix is a type of data that is often used in R programming. This is used primarily in the input to obtain word frequencies. The corpus variable can be type cast into a matrix of the text name. We also generate word cloud for the selected data set for better visualization. In the cloud, not all the words are shown, but words with more frequencies are identified and shown as a word cloud.

A. Result and Discussion

A total of 1,305,000 tweets from 11 infected countries were collected during the study period. The results of the research are discussed in two parts.

In the first part, the sentiments of the tweets from all the 11 infected countries were discussed. Fig. 2, shows the sentiments of tweets by the people of the eleven infected countries for which the study was conducted. It is evident from the Fig. 3 that the tweets from almost infected countries had positive sentiments. In countries, like USA and Chile had almost a balance between positive and negative sentiments, whereas, Brazil had 55% positive sentiments and 46% negative sentiments. India was followed by Peru, Spain and UK where 55% of the peoples were tweeting with a positive attitude while 45% with negative attitude. In Russia 51% of the peoples were tweeting with a positive attitude while 49% with negative attitude. In South Africa, Mexico and Oman has 54%, 49% and 57% of the peoples were tweeting with a positive attitude while 46%, 51% and 43% with negative attitude. In Oman maximum people expressed positive attitude since the recovery rate is high.

In the second phase, the emotions associated with the collected tweets were analyzed. In this process, it was observed that almost all countries had the highest number of tweets with more trust, since the recovery rate is high in almost all infected countries. Almost all countries had the highest number of tweets with fear initially due to more number of people infected with COVID-19. The country with most number of tweets associated with Fear and Sadness was India while the anticipation quotient was highest in the tweets from Spain. There were also a good number of tweets with the emotion of Trust which attributed to the invention of medicine is going on different countries. Sentiments are extracted as outputs from functions of the syuzhet package from confidential data and the sentiment approaches categorize the text with its corresponding sentiment meaning. A line chart in Fig. 3 shows the feelings examined from the classified texts.

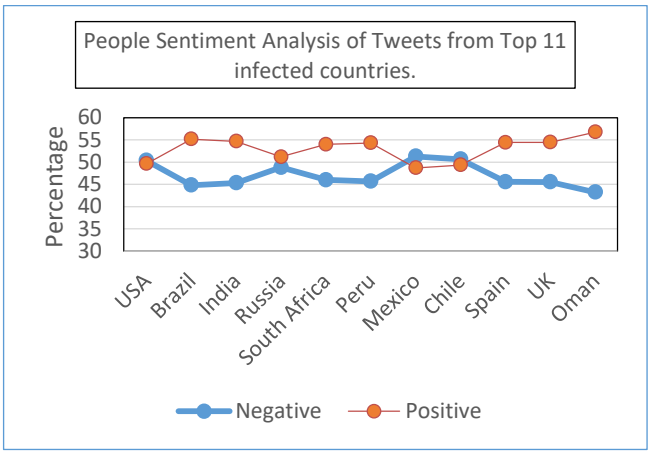


Fig. 2. Sentiment Analysis of Tweets among different Infected Countries.

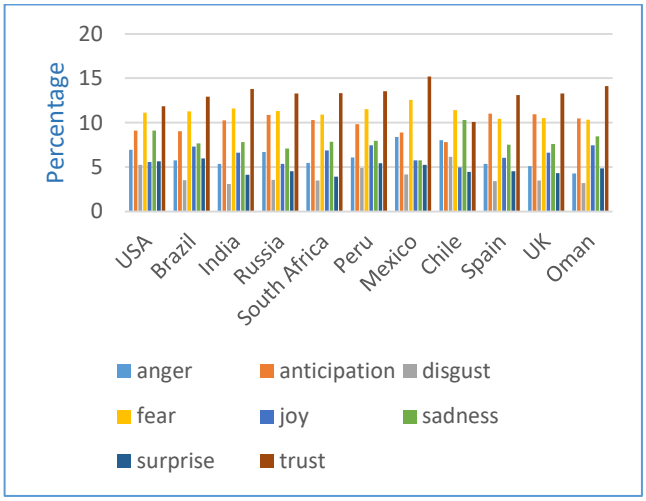


Fig. 3. Emotion Analysis of the Tweets from Infected Country.

After this, the tweets were organized into word clouds to analyze what words have been frequently used by the twitter users of different countries and also what emotions were behind these words. As it can be seen from Fig. 4 to Fig. 14, words like Pandemic, COVID, Virus, Hospitals, Health, Fight, Stay, Safe, Help, Emergency and Death were very frequently used by the users of almost from all countries.

In USA, the words such as trump, Pandemic and Death associated with Surprise, Sadness and Disgust, were the sentiments mostly used in the tweets. In Brazil, words like Pandemic, COVID, Fight, Death and Pandemic were mostly used with the emotions of sadness, anger and disgust respectively. People of India and South Africa used the words in tweets such as Pandemic, Hospital, Death and Fight associated with emotions of sadness, disgust and anger. People in Russia tweeted using the words Pandemic, trump and protest with emotions of sadness, surprise and disgust. People in Peru used the words such as pandemic, stay home and distance with the emotions of sadness and joy. This was evident because the people were now aware of COVID-19 precautions. Mexico had used the words such as covid, help and health with emotions of joy. In Spain, mostly people used the words in tweets such as pandemic, covid and death with emotions sadness, fear and disgust. Citizens of UK were using words like Pandemic, Government and death very frequently which were associated with emotions of sadness, fear and disgust. Similarly, people across Oman used words like Pandemic, covid and death which were used to emote the feeling of sadness and disgust.

However, across all the tweets analysed from eleven infected countries, there was a very good amount of mentions of a political personality. The name of US President, Donald Trump appeared consistently in many tweets across all countries. These mentions were mostly associated with the emotion of surprise.

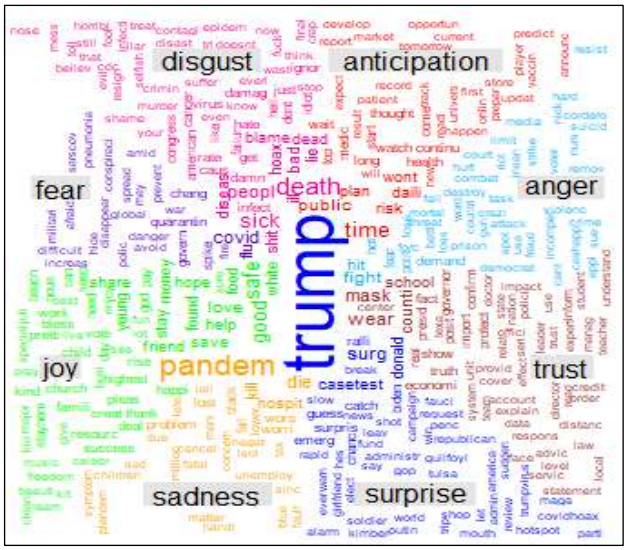


Fig. 4. Word Cloud Tweets from USA.

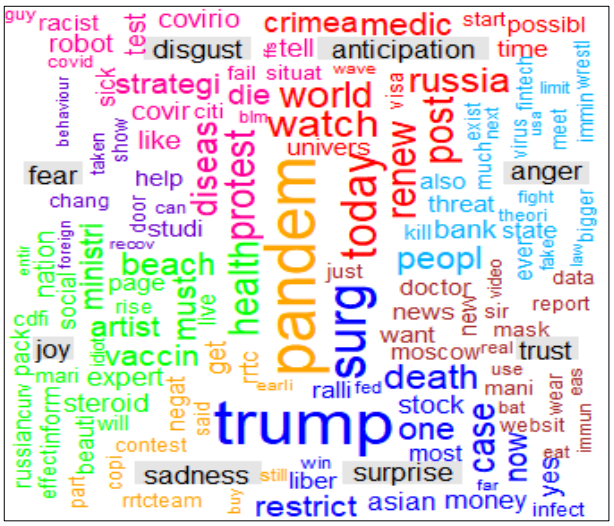


Fig. 5. Word Cloud Tweets from Russia.

Fig. 6. Word Cloud Tweets from Brazil.

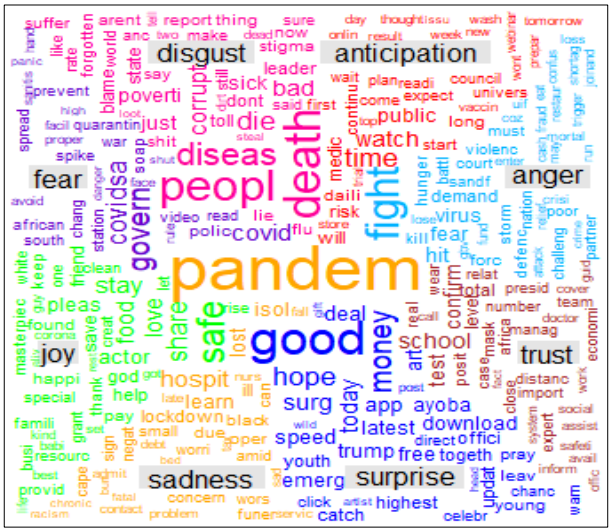


Fig. 7. Word Cloud Tweets from South Africa.

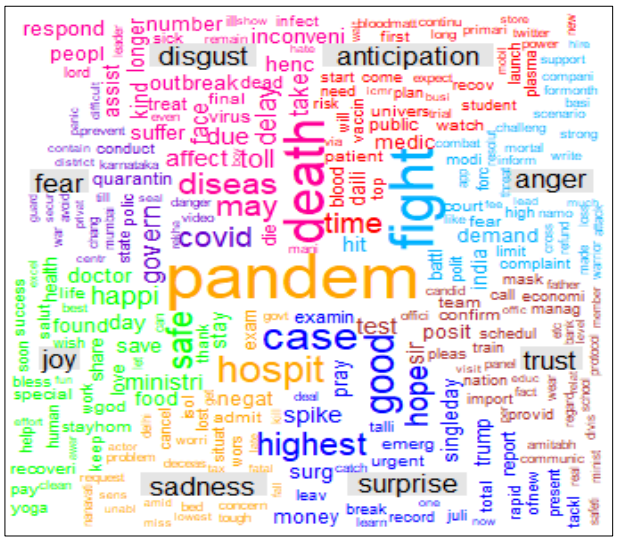


Fig. 8. Word Cloud Tweets from India.

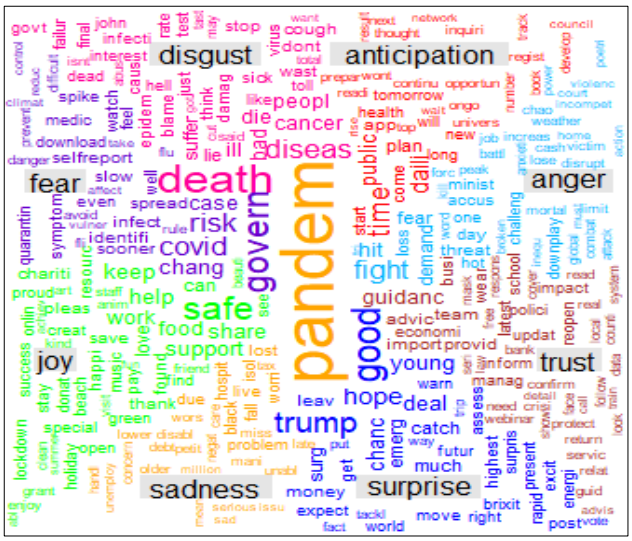


Fig. 11. Word Cloud Tweets from UK.

Fig. 9. Word Cloud Tweets from Peru.

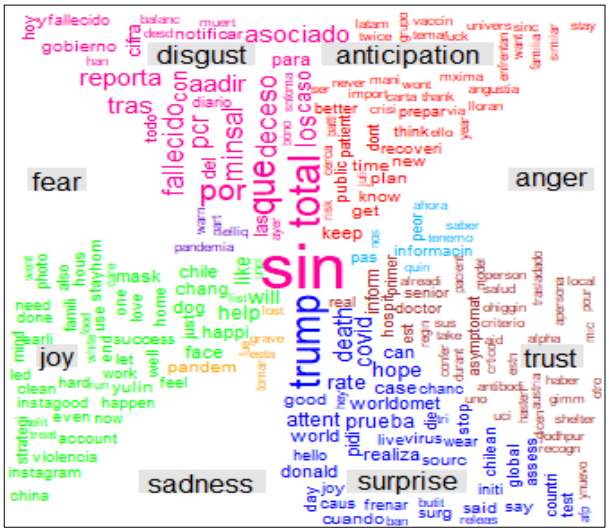


Fig. 12. Word Cloud Tweets from Chile.

Fig. 10. Word Cloud Tweets from Mexico.

Fig. 13. Word Cloud Tweets from Spain.

Fig. 14. Word Cloud Tweets from Oman.

B. Limitations

It is important to mention that there were some limitations to this study. First, we used a comprehensive hashtag list, which was limited by the knowledge of the authors about trending hashtags and imagination. We have overlooked alternate words or incorrect spellings and implemented with biased selection in the tweets that we have reviewed. Secondly, we addressed tweets in English; thus, our findings may not be generalized one compared to some countries where English is not the main language.

VI. Conclusion

This research work aimed at analyzing the sentiments and emotions of the people during the pandemic COVID-19 have been successfully conducted. To maintain the credibility of data and also the ease of extracting tweets of users, the Twitter platform has been chosen for the study. To have reasonable and good useful data for the study, the twitter users of top 10 infected countries along with one country chosen from the Gulf region for this work. During the study, it was observed that almost all countries were tweeting about COVID19 with a positive sentiments, since all those people become habitual to COVID-19 and also the recovery rate has been improved over the time. Similarly, while analysing the word clouds of different countries, it was concluded that people are tweeting words like Pandemic, COVID, Virus, Hospitals, Health, Fight, Stay, Safe, Help, Emergency, Death and Masks with different emotions. This study provided a good analysis on sentiments and mind sets of people on Covid-19 and enabled us to understand that almost the same level of thinking of people all over the world. This research can be used for future works to examine the shifting emotions and feelings of individuals from these nations and to check if there are noticeable changes over time in them.

References

[1] https://www.worldometers.info/coronavirus/, accessed on July 11, 2020.

[2] Rosenberg, H., Syed, S., & Rezaie, S. (2020). The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. Canadian Journal of Emergency Medicine, 1-4.

[3] Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. American journal of infection control, 38(3), 182-188.

[4] Guy, M., Earle, P., Ostrum, C., Gruchalla, K., & Horvath, S. (2010, May). Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In International symposium on intelligent data analysis (pp. 42-53). Springer, Berlin, Heidelberg.

[5] Cheong, M., & Lee, V. C. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. Information Systems Frontiers, 13(1), 45-59.

[6] Rufai, S. R., & Bunce, C. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. Journal of Public Health.

[7] Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. Journal of medical Internet research, 22(4), e19016.

[8] Ahmed W, Vidal-Alaball J, Downing J, López Seguí, F. COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. Journal of Medical Internet Research 2020;22(5):e19458. doi:10.2196/19458.

[9] Alvarez-Risco A, Mejia CR, Delgado-Zegarra J, Del-Aguila-Arcentales S, Arce-Esquivel AA, Valladares-Garrido MJ, . . . Yáñez, JA. The Peru Approach against the COVID-19 Infodemic: Insights and Strategies. The American Journal of Tropical Medicine and Hygiene 2020 June 04:1-4. doi:10.4269/ajtmh.20-0536.

[10] Budhwani H, Sun R. Creating COVID-19 Stigma by Referencing the Novel Coronavirus as the "Chinese virus" on Twitter: Quantitative Analysis of Social Media Data. Journal of Medical Internet Research 2020;22(5):e19301.

[11] Chen E, Lerman K, Ferrara E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Health and Surveillance 2020;6(2):e19273. doi:10.2196/19273.

[12] Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, . . . Cuomo R. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. JMIR Public Health and Surveillance 2020;6(2): e19509. doi:10.2196/19509.

[13] C.Kaur and A. Sharma, "Twitter Sentiment Analysis on Coronavirus using Textblob," EasyChair2516 -2314, 2020.

[14] M. Ra, B. Ab, and S. Kc, "COVID 19 Outbreak: Tweet based Analysis and Visualization towards the Influence of Coronavirus in the World.".

[15] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. J. m. Lehmann, "An" Infodemic": Leveraging High Volume Twitter Data to Understand Public Sentiment for the COVID 19 Outbreak," 2020.

[16] Samuel, J., Ali, G. G., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Information, 11(6), 314. doi:10.3390/info11060314.

[17] Carvalho, J.P.; Rosa, H.; Brogueira, G.; Batista, F. MISNIS: An intelligent platform for twitter topic mining. Expert Syst. Appl. 2017, 89, 374–388.

[18] De Choudhury, M.; Counts, S.; Horvitz, E. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 3267–3276.

[19] Samuel, J.; Holowczak, R.; Benbunan-Fich, R.; Levine, I. Automating Discovery of Dominance in Synchronous Computer-Mediated Communication. In Proceedings of the 2014 47th IEEE Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 6–9 January 2014; pp. 1804–1812.

[20] ansen, B.J., Zhang, M., Sobel, K., Chowdury, A., 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American societyfor information science and technology 60, 2169–2188.

[21] Wilson, T., Wiebe, J., & Hoffmann, P. Recognizing contextual polarity in phraselevel sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics(2005, October).

[22] Liu, B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167(2012).

[23] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30- 38). Association for Computational Linguistics(2011, June).

[24] Sulthana, A. R., Jaithunbi, A. K., & Ramesh, L. S. (2018). Sentiment analysis in twitter data using data analytic techniques for predictive modelling. Journal of Physics: Conference Series, 1000, 012130. doi:10.1088/1742-6596/1000/1/012130.

[25] Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. American journal of infection control, 38(3), 182-188.

[26] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." In Cognitive Informatics and Soft Computing, pp. 639-647. Springer, Singapore, 2019.

[27] Uma Ramya V, Thirupathi Rao K. Sentiment Analysis of Movie Review using Machine Learning Techniques. International Journal of Engineering & Technology. 2018;7(2.7):676.

[28] M A, Ravikumar P. Survey: Twitter data Analysis using Opinion Mining. International Journal of Computer Applications. 2015;128(5):34-36.

[29] Maindola, Pallavi, Neetu Singhal, and Akash D. Dubey. "Sentiment Analysis of Digital Wallets and UPI Systems in India Post Demonetization Using IBM Watson." In 2018 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6. IEEE, 2018.

[30] Wang Y, Taylor J. Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake. Natural Hazards. 2018;92(2):907-925.

[31] Chorianopoulos K, Talvis K. Flutrack.org: Open-source and linked data for epidemiology. Health Informatics J 2016; 22:962–974.

[32] Signorini A, Segre AM, Polgreen PM. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS ONE 2011; 6:e19467.

[33] Househ M. Communicating Ebola through social media and electronic news media outlets: A cross-sectional study. Health Informatics J 2016; 22:470–478.

[34] Matthew L. Jockers. Syuzhet: Extract Sentiment and Plot Arcs from Text. 2015. Available at: https://github.com/mjockers/syuzhet. Accessed 30 January 2020.

[35] Colneric N, Demsar J. Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. IEEE Trans Affective Comput 2019; :1–1.

[36] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research 2003; 3:993–1022.

[37] Mohammad S, Turney P. Crowdsourcing A Word-Emotion Association Lexicon. Computational Intelligence. 2012;29(3):436-465.

[38] Plutchik R. The Nature of Emotions. American Scientist. 2001;89(4):344.

[39] Jockers M. Introduction to the Syuzhet Package [Internet]. Cran.r-project.org. 2017. Available at: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html.

[40] Problems with the Syuzhet Package [Internet]. Anglophile in Academia: Annie Swafford's Blog. 2015 [cited 2020 Apr 3]. Available from: https://annieswafford.wordpress.com/2015/03/02/syuzhet/.

[41] Su, Yuanhang, Ruiyuan Lin, and C-C. Jay Kuo. "Tree-structured multi-stage principal component analysis (TMPCA): Theory and applications." Expert Systems with Applications 118 (2019): 355-364.

[42] Su, Yuanhang, et al. "Unsupervised multi-modal neural machine translation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[43] Su, Yuanhang, Yuzhong Huang, and C-C. Jay Kuo. "Efficient text classification using tree-structured multi-linear principal component analysis." 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018.

[44] Su, Yuanhang, and C-C. Jay Kuo. "On extended long short-term memory and dependent bidirectional recurrent neural network." Neurocomputing 356 (2019): 151-161.

The Enrichment of Texture Information to Improve Optical Flow for Silhouette Image

Bedy Purnama[1]*, Mera Kartika Delimayanti[2]
Kunti Robiatul Mahmudah[3], Fatma Indriani[4], Mamoru Kubo[5], Kenji Satou[6]
Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa 9201192, Japan[1, 2, 3, 4]
Telkom of Computing, Telkom University, Bandung 40257, Indonesia[1]
Department of Computer and Informatics Engineering, Politeknik Negeri Jakarta, Depok 16425, Indonesia[2]
Institute of Science and Engineering, Kanazawa University, Kanazawa 9201192, Japan[5, 6]

***Abstract*—Recent advances in computer vision with machine learning enabled detection, tracking, and behavior analysis of moving objects in video data. Optical flow is fundamental information for such computations. Therefore, accurate algorithm to correctly calculate it has been desired long time. In this study, it was focused on the problem that silhouette data has edge information but does not have texture information. Since popular algorithms for optical flow calculation do not work well on the problem, a method was proposed in this study. It artificially enriches the texture information of silhouette images by drawing shrunk edge on the inside of it with a different color. By the additional texture information, it was expected to give a clue of calculating better optical flows to popular optical flow calculation algorithms. Through the experiments using 10 videos of animals from the DAVIS 2016 dataset and TV-L1 algorithm for dense optical flow calculation, two values of errors (MEPE and AAE) were evaluated and it was revealed that the proposed method improved the performance of optical flow calculation for various videos. In addition, some relationships among the size of shrunk edge and the type and the speed of movement were suggested from the experimental results.**

Keywords*—Optical flow; silhouette image; artificial increase of texture information

I. INTRODUCTION

Today, a number of cameras are installed in various devices such as smart phones, PCs, security devices, etc. They generate a large amount of still images and videos every day resulting the demand to analyze images and videos by computers has been growing rapidly. In the research field of computer vision, machine learning methods including deep learning have been actively studied for the purpose of achieving better performance. In addition, some studies which have been done by employing the combination of these methods significantly improved the performance [1].

One of the rapidly developing studies of computer vision and deep learning are the detection of moving objects [2] [3], where the most fundamental information is optical flow [4] [5] [6] that indicates corresponding points in two images. In the case of video data, typically the two images are i^{th} and $(i+1)^{th}$ video frames. It is useful for detecting and tracking moving objects in a video.

Among many studies in computer vision, this study focused on the analysis of animal behavior. Similar to the analysis of human behavior, analysis of animal behavior brings many benefits to bionomics, ethology, medical science, fishery, pet industry, etc. It has been studied a long time, mainly in the level of animal's location tracking (coarse-grain level). There are some studies on applying optical flow to detect animal's presence such as outdoor animal detection using Neural Network [7], elephant detection by using color model [8], pig and deer detection by using CNN [9] [10]. On the other hand, analysis of animal's action or motion (fine-grain level) has also attracted interests of researchers in the interdisciplinary field between life science and computer vision.

For the analysis of animal's action or motion, dense optical flow (i.e., optical flow for every pixel of object) should be calculated accurately. The basis of optical flow was developed by Horn and Schunck [11]. In general, optical flow is calculated based on two kinds of information, that is, edges and texture of objects. However, sufficiently rich texture information of animal is not always available. For example, a unicolored animal like a black cat may provide clear edge information, but does not provide any texture information on its body. Moreover, even if an animal has clear texture information on its body surface, it is considered as a silhouette in a backlit video. The previous research focused on how the accuracy of optical flow for videos containing silhouette images of animals can be improved [12]. Since no texture information is available for a silhouette image, popular methods of optical flow calculation provided returned wrong or missing optical flow, especially for the center area of silhouette images (in contrast, optical flows near to the edge are relatively accurate since edge information is available also for silhouette images). To solve the problem, some areas to be improved and additionally calculated by an inpainting algorithm using perspective transformation have been detected in this method in order to solve the problem. However, this method has not been applied to more complex silhouettes.

A new method for improving the accuracy of optical flow for silhouette images is proposed in this paper. In short, the texture information of a silhouette image has been enriched by drawing shrunk edge on the inside of it with a different color (it is called object-in-object). In this research, it is assumed that the object (animal) has already been separated from the background.

*Corresponding Author

To estimate the effect of the proposed method, some videos of animals from a publicly available video dataset were used to estimate the effect of the method. In the selected videos, only one animal is seen. After segmenting an animal from the background of each video, a video which consists of silhouette images was generated. Through the comparison of optical flows for original and silhouette video, it was shown the accuracy of optical flow for silhouette images can be improved in this method. The main contributions of this research are as follows.

- Focusing on silhouette images, seeking novelty to improve the accuracy of existing dense optical flow methods.
- The addition of texture (object-in-object) can increase the accuracy of optical flow for silhouette images.
- The improvement was estimated by using a publicly available video dataset.

II. Background

The research regarding the use of feature texture in order to improve the optical flow estimation performance has been conducted by the researchers. It was initiated by Arredondo [13] using differential method to estimate optical flow from feature texture. He used mathematical approach for this research.

On the other hand, the empirical approach was used by Andalibi [14]. Andalibi added the static texture in images with poor texture. The addition of texture mask, where the image area had its optical flow component approaching to zero, did not significantly improve the optical flow estimation. Nevertheless, this was an initial approach to solve the problems on poor texture in optical flow.

In the researches [13] and [14], there are some chances to improve the optical flow estimation for silhouette image sequences cases. On silhouette image sequences, there was almost nothing to find out the feature texture. It was expected that with the addition of texture information to silhouette image sequences, it improved the optical flow performance.

III. Materials and Methods

A. Silhouette Image and Video

The previous research [12] focused on animal silhouette images animated by a program (i.e. rotation). The uniqueness of the silhouette animal was that it only had edge information with no texture information. In real conditions, it is believed that animals can be found if they are walking at dusk or they have unicolor, such as black horses. To evaluate the new method proposed in this paper, the DAVIS 2016 dataset [15] was used in this study. Since important objects in each video in the dataset have already been segmented, it assumes that the animal in a video has already been separated from the background. Practically, the result of segmentation is given as a mask image to the original image of each frame of a video. The region of an animal indicated by the mask image is also referred as a region of Interest (RoI). By applying the mask image to the original image, the silhouette image is generated, where the RoI is black and the background is white.

The use of image silhouettes from animal videos revealed a variety of natural animal movements, where each body part of an animal is freely moving or not moving. In case of a simple movement such as walking, only some parts of body are moving. In contrast, almost every body parts are moving in a more complex movement like jumping or flying. From the dataset DAVIS 2016, which contains 50 videos with 480p resolution, 10 videos shown in Table I were chosen in accordance with the criteria that a unicolored animal is taking natural motion in the video. Walk movement shows that animals walk on their feet. Move forward movement indicates that the animal moved without walking. Deformation movement indicates a change in the direction of the animal's orientation. Slow indicates camera / animal movement tends to be slow while fast refers to fast camera / animal movement.

B. Optical Flow

The basic concept of optical flow was initiated by Horn and Schunck [11]. The formulation in (1) is a general equation for the optical flow. There are two solutions that must be resolved, namely, data term and smoothness term. There are traditional developments for optical flow algorithms such as Brox [16], Black [17], Lucas-Kanade [18], TV-L1[19]. There are also the studies that use deep learning such as Flownet [20], DeepFlow [21], EpicFlow [22].

$$E(u,v) = \int_{\Omega} \underbrace{(I_x u + I_y v + I_t)^2}_{data\ term} + \underbrace{\lambda(|\nabla u|^2 + |\nabla v|^2)}_{smoothness\ term} dxdy \qquad (1)$$

To form the ground truth of this study, DeepFlow was used [21]. RGB images from the dataset were used to form. This DeepFlow was divided into two processes, namely, feature search using DeepMatching [23] to handle large displacements. The next process was DeepFlow itself, which used deep learning. The output of the DeepFlow is .flo file format that was adjusted to the Middlebury dataset standard [24] [25].

To generate optical flow from silhouette images, the TV-L1[26] [27] was used since this method is the most commonly used as a baseline for optical flow. TV-L1 is based on the coarse method to fine framework numerically, while DeepFlow is based on coarse to fine in the deep learning framework.

TABLE I. Dataset used in this Study

Data Name	Number of Frames	Type of Movements
Bear	82	Walk, slow
Blackswan	50	Move forward, slow
Camel	90	Walk, deformation, slow
Cows	104	Walk, deformation, slow
Dog	60	Walk, deformation, fast
Elephant	80	Walk, deformation, slow
Flamingo	80	Walk, deformation, slow
Goat	90	Walk, deformation, fast
Mallard-fly	70	Walk, fly, high deformation
Mallard-water	80	Move forward, deformation, slow

C. Framework

Fig. 1 illustrates the framework of the experiment in this study. Fig. 2 is the initial stage of the framework. It starts with the formation of Ground Truth from two consecutive RGB JPEG images with a resolution of 480p for each data. DeepFlow [21] was used and the output of this process (i.e. the Ground Truth) is named optical flow t.

The next process described in Fig. 3 is the novelty of this study. From the DAVIS 2016 dataset, annotations (i.e., mask images) with 480p resolution have been provided. In this process, for the enrichment of texture information, a shrunk edge line with white color was drawn at the inside of RoI with black color. From here, a shrunk edge line is called as an artificial texture. This artificial texture was located inside the RoI, where the RoI centroid coincided with the centroid of the artificial texture. In the experiment, the magnification sizes from 10, 20, 30, 40, 50, 60, 70, 80, and 90 percentages were tried.

The silhouette image that has been enriched with an artificial texture was named silhouette image t', which was the output from the silhouette image input t. As shown in Fig. 4, silhouette image t' and silhouette image t' +1 were the input for TV-L1 to form a new optical flow t'.

The final stage was performance evaluation. From Fig. 5, it can be observed that mean endpoint error (MEPE) was used and average angular error (AAE) [24] for evaluating the accuracy of optical flow calculated from silhouette images. MEPE was used to calculate the average magnitude deviation of the vector flow from the estimated optical flow to the ground truth. MEPE was calculated by (2). Meanwhile, AAE was to calculate the average angular deviation of vector flow from the estimated optical flow towards the ground truth. AAE was calculated using (3).

$$EPE = \sqrt{(u - u_G)^2 + (v - v_G)^2} \quad (2)$$

$$AAE = cos^{-1}\left(\frac{1.0 + u \times u_G + v \times v_G}{\sqrt{1.0 + u^2 + v^2}\sqrt{1.0 + {u_G}^2 + {v_G}^2}}\right) \quad (3)$$

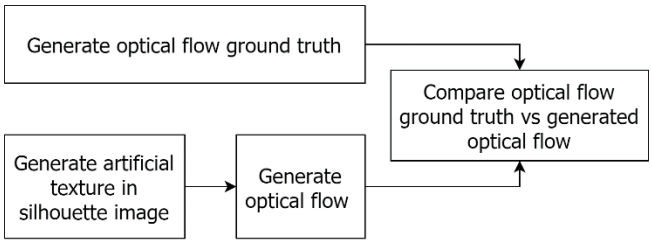


Fig. 1. The Framework of the Experiment.

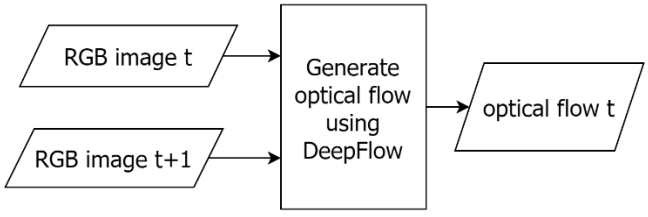


Fig. 2. Generated Ground Truth of Optical Flow.

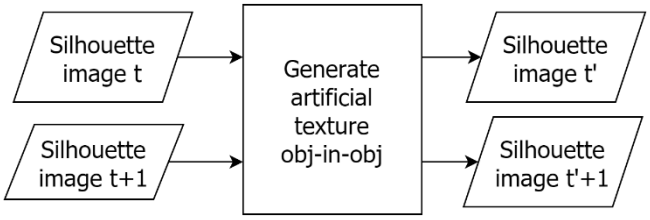


Fig. 3. Generate Artificial Texture.

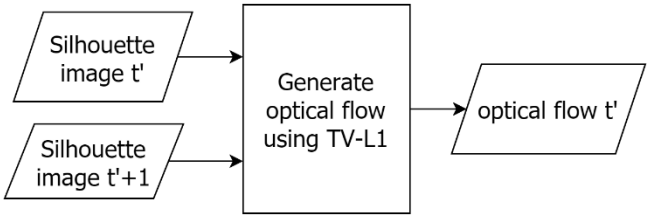


Fig. 4. Generated Optical Flow from Sequence of Silhouette Artificial Texture.

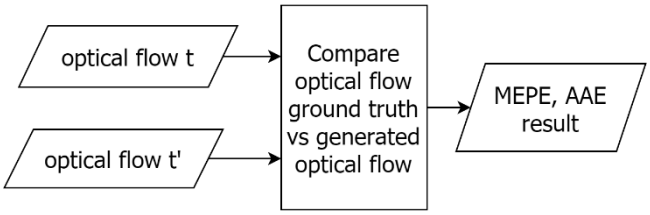


Fig. 5. Compare Optical Flow Performance.

IV. Result

In this study, the supporting applications used were:

- OpenCV version 4.11 with contrib which provided TV-L1.
- DeepMatching version 1.2.
- DeepFlow version 2.

The experimental results can be seen in Fig. 6 and Fig. 7. Fig. 6 shows an example of ground truth, where two RGB image sequences were processed by DeepFlow algorithm to generate ground truth optical flow. Using the same example, Fig. 7 shows the silhouette images with artificially enriched texture. The first column of it shows the percentage of artificial texture which is gradually added by 10% until it reaches the maximum of 90%. The second and third columns show the silhouette images corresponding to t^{th} and $(t+1)^{th}$ frames respectively. The last column shows the optical flow calculated by TV-L1 algorithm with two silhouette images in the same row as input. In the row named "without" in Fig. 7, it is clearly shown that an almost white area is observed at the center of RoI. It indicates that TV-L1 does not work well for silhouette images without enrichment. Moreover, it is demonstrated that the enrichment of texture might solve this problem.

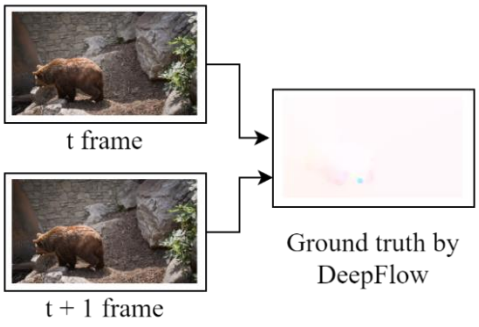


Fig. 6. Two Consecutive Original Images and its Optical Flow (Ground Truth) Calculated by DeepFlow.

Table II shows the MEPE calculated from optical flow of silhouette images against optical flow ground truth. Starting from no increments, increments of 10 percent to 90 percent of the original size. Meanwhile, Table III shows the AEE for the same case.

In this study, the results of optical flow were evaluated with two measurements, namely, MEPE and AAE were evaluated. Both were achieved by comparing the optical flow of the experimental results with ground truth. To determine the quality of MEPE and AAE, it can be observed from the value, the smaller the value the better the performance.

Table II revealed that there were six animal data (Camel, Cows, Dog, Elephant, Flamingo' and Mallard-fly) that achieved the lowest MEPE value with the formation of 90% of magnification artificial texture. In contrast, in the Bear and Blackswan data, the lowest MEPE value was in the formation of 80% of magnification artificial texture. What was different from the others were that Goat and Mallard-water each achieved the lowest MEPE value at the formation of 50% and 70% of magnification artificial texture.

Table III revealed that there were five animal data (Camel, Cows, Dog, Elephant, and Flamingo) which achieved the lowest AAE value in the formation of 90% of magnification artificial texture. In contrast, in the Bear and Blackswan data, the lowest AAE value was in the formation of 80% of magnification artificial texture. What was different from the others were that Goat, Mallard-fly, and Mallard-water, where each achieved the lowest AAE values at the formation of 50%, 60%, and 70% of magnification artificial texture.

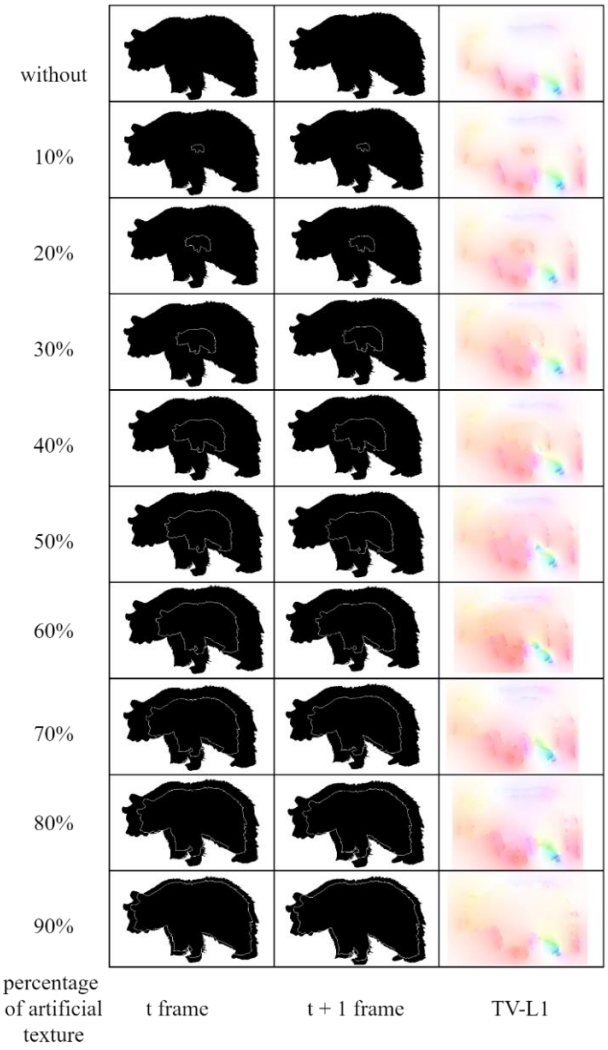


Fig. 7. RoI with Artificially Enriched Texture and its Optical Flow by TV-L1.

TABLE II. MEPE OF GROUND TRUTH BY TV-L1 ARTIFICIAL TEXTURE

	Percentage of magnification artificial texture									
	w/o	***10***	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***90***
Bear	1,136	1,098	1,069	1,046	1,028	1,018	1,006	0,999	**0,997**	1,000
Blackswan	1,844	1,793	1,762	1,737	1,713	1,693	1,687	1,664	**1,641**	1,650
Camel	1,432	1,392	1,361	1,333	1,320	1,308	1,299	1,288	1,270	**1,260**
Cows	1,489	1,450	1,415	1,381	1,364	1,351	1,343	1,339	1,338	**1,333**
Dog	6,133	6,110	6,093	6,070	6,067	6,065	6,092	6,125	6,125	**5,940**
Elephant	1,425	1,417	1,409	1,400	1,386	1,373	1,368	1,359	1,351	**1,344**
Falmingo	1,717	1,708	1,697	1,686	1,675	1,668	1,661	1,655	1,647	**1,639**
Goat	3,897	3,865	3,834	3,800	3,780	**3,775**	3,787	3,809	3,836	3,841
Mallard- fly	5,447	5,436	5,427	5,418	5,412	5,412	5,416	5,421	5,423	**5,411**
Mallard -water	2,109	2,082	2,060	2,042	2,029	2,020	2,017	**2,015**	2,024	2,029

TABLE III. AAE OF GROUND TRUTH BY TV-L1 ARTIFICIAL TEXTURE

	Percentage of magnification artificial texture									
	w/o	***10***	***20***	***30***	***40***	***50***	***60***	***70***	***80***	***90***
Bear	36,813	35,541	34,519	33,630	32,863	32,419	31,882	31,583	**31,504**	31,553
Blackswan	46,164	44,334	43,235	42,316	41,425	40,635	40,352	39,422	**38,470**	38,688
Camel	36,734	35,496	34,435	33,438	32,964	32,544	32,233	31,821	31,261	**30,760**
Cows	42,889	41,395	40,014	38,635	37,937	37,397	37,081	36,842	36,722	**36,383**
Dog	51,337	50,488	49,627	48,741	48,263	47,744	47,897	48,167	48,044	**44,988**
Elephant	37,370	37,000	36,604	36,190	35,591	35,051	34,877	34,497	34,065	**33,786**
Falmingo	43,213	42,869	42,445	42,059	41,625	41,330	41,060	40,797	40,395	**39,866**
Goat	56,694	55,754	54,742	53,680	52,859	**52,388**	52,446	52,823	53,271	53,352
Mallard- fly	56,054	55,621	55,159	54,669	54,270	54,070	**53,946**	54,030	54,261	54,406
Mallard -water	56,054	55,153	54,458	53,873	53,478	53,160	52,979	**52,794**	52,930	53,018

V. DISCUSSION AND CONCLUSIONS

All the experiments revealed that the addition of an artificial texture to the silhouette image can improve performance. Thus, the proposed method worked well for realistic silhouette images generated from DAVIS 2016 dataset. Moreover, it can be observed that there was no MEPE or the smallest AAE in the w/o column (without adding an artificial texture).

In Table II and Table III, from five animal data (Camel, Cows, Dog, Elephant, and Flamingo) the lowest MEPE and AAE were in subcolumn 90. If it is connected with Table I which revealed the motion type, it can be noticed that five of them had the same type of motion type on walk, deformation and slow. Only Dog was the fast type. This experiment revealed that for the case of relatively slow object movement also little object deformation and adding an artificial texture that is very close to the original texture can improve performance.

For the case of Bear and Blackswan data, the smallest MEPE and AAE values were in the 80 subcolumn, where both motion types were slow. For Bear was the walk movement while Blackswan was the moves forward movement. It was revealed that in the case of slow moving objects without deformation, the addition of artificial texture was still required to improve performance. However, it was close enough to the original texture at about 80% of the original size.

The rest of the three animal data were slightly different from the others. Mallard-fly achieved the lowest MEPE in subcolumn 90, but it achieved the lowest AAE in sub-column 60 where the types of movement were walk, high deformation, and fly. There was an inconsistency between MEPE and AAE, which were usually in the same subcolumn. This was due to a large deformation change, from walking to flying. As for Mallard-water, the lowest MEPE and AAE values were in the 70 subcolumn. In terms of type of movement, it was almost similar to Blackswan's, namely, move forward and slow movements. However, there was additional deformation of the object's motion orientation. From the Mallard-water experiment, it can be observed that where the movement was slow and where the deformation was progressive, it was still necessary to add an artificial texture, but it was a bit far from the original. The last one that was the most different from the other data were the goat data. The lowest MEPE and AAE values fell in sub column 50, with the type of walk movement, deformation and fast. What distinguishes it from the others is primarily the deformation of the object. This was due to the movement of the fur. Afterward, the camera movement was great for following objects. From this experiment, it was revealed that with large object deformation and camera movement, only 50% of the artificial texture was needed to improve performance.

Improved silhouette image performance by adding artificial texture (object-in-object method) can be categorized as follows:

- The addition of an artificial texture as much as 90% of the original texture, for slowly moving objects and simple deformations.
- The addition of 70% or 80% artificial texture of the original texture, for fast moving objects and simple deformations.
- The addition of artificial texture 60% 50% of the original texture, for fast moving objects and large deformations.

This study attempted to compare the ground truth from real video in the form of RGB images and optical flow which was formed from modified image silhouette. The contribution of the proposed method to modify the silhouette by adding an artificial texture can actually improve its performance. The performance measurements used were MEPE and AAE. The limitation of this study was it was only for a single segmented object. For the case of multiple objects, the object of occlusion was not included in this study and it can be studied for further research.

ACKNOWLEDGMENT

The first and second authors would like to gratefully acknowledge to BUDI-LN scholarship from Indonesia Endowment Fund for Education (LPDP), Ministry of Education and Culture, Republic of Indonesia (KEMENDIKBUD), and Ministry of Research and Technology of Republic Indonesia (KEMENRISTEK). In this research, the super-computing resource was provided by Human

Genome Center, the Institute of Medical Science, the University of Tokyo. Additional computation time was provided by the super computer system in Research Organization of Information and Systems (ROIS), National Institute of Genetics (NIG).

REFERENCES

[1] N. O'Mahony et al., "Deep Learning vs. Traditional Computer Vision," in Advances in Computer Vision, vol. 943, K. Arai and S. Kapoor, Eds. Cham: Springer International Publishing, 2020, pp. 128–144.

[2] M.-N. Chapel and T. Bouwmans, "Moving objects detection with a moving camera: A comprehensive review," Comput. Sci. Rev., vol. 38, p. 100310, Nov. 2020, doi: 10.1016/j.cosrev.2020.100310.

[3] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object Detection With Deep Learning: A Review," IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 11, pp. 3212–3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[4] Y. Zhang, J. Zheng, C. Zhang, and B. Li, "An effective motion object detection method using optical flow estimation under a moving camera," J. Vis. Commun. Image Represent., vol. 55, pp. 215–228, Aug. 2018, doi: 10.1016/j.jvcir.2018.06.006.

[5] D. H. Ye, J. Li, Q. Chen, J. Wachs, and C. Bouman, "Deep Learning for Moving Object Detection and Tracking from a Single Camera in Unmanned Aerial Vehicles (UAVs)," Electron. Imaging, vol. 2018, no. 10, pp. 466-1-466–6, Jan. 2018, doi: 10.2352/ISSN.2470-1173.2018.10.IMAWM-466.

[6] J. Huang, W. Zou, J. Zhu, and Z. Zhu, "Optical Flow Based Real-time Moving Object Detection in Unconstrained Scenes," ArXiv180704890 Cs, Jul. 2018, Accessed: Jan. 03, 2021. [Online]. Available: http://arxiv.org/abs/1807.04890.

[7] M. Bonneau, J.-A. Vayssade, W. Troupe, and R. Arquet, "Outdoor animal tracking combining neural network and time-lapse cameras," Comput. Electron. Agric., vol. 168, p. 105150, Jan. 2020, doi: 10.1016/j.compag.2019.105150.

[8] M. Zeppelzauer, "Automated detection of elephants in wildlife video," EURASIP J. Image Video Process., vol. 2013, no. 1, p. 46, Dec. 2013, doi: 10.1186/1687-5281-2013-46.

[9] L. Zhang, H. Gray, X. Ye, L. Collins, and N. Allinson, "Automatic Individual Pig Detection and Tracking in Pig Farms," Sensors, vol. 19, no. 5, p. 1188, Mar. 2019, doi: 10.3390/s19051188.

[10] W. J. Hans, V. Sherlin, and N. Venkateswaran, "On-Road Deer Detection for Advanced Driver Assistance using Convolutional Neural Network," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, 2020, doi: 10.14569/IJACSA.2020.0110499.

[11] B. K. P. Horn and B. G. Schunck, "Determining optical flow," Artif. Intell., vol. 17, no. 1–3, pp. 185–203, Aug. 1981, doi: 10.1016/0004-3702(81)90024-2.

[12] B. Purnama et al., "On the Problem about Optical Flow Prediction for Silhouette Image:," 2019, pp. 289–293, doi: 10.5220/0007572302890293.

[13] M. A. Arredondo, K. Lebart, and D. Lane, "Optical flow using textures," Pattern Recognit. Lett., vol. 25, no. 4, pp. 449–457, Mar. 2004, doi: 10.1016/j.patrec.2003.11.007.

[14] M. Andalibi, Lawrence. L. Hoberock, and H. Mohamadipanah, "Effects of texture addition on optical flow performance in images with poor texture," Image Vis. Comput., vol. 40, pp. 1–15, Aug. 2015, doi: 10.1016/j.imavis.2015.04.008.

[15] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, Jun. 2016, pp. 724–732, doi: 10.1109/CVPR.2016.85.

[16] T. Brox, "Optical Flow: Traditional Approaches," in Computer Vision, Cham: Springer International Publishing, 2020, pp. 1–5.

[17] M. J. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields," Comput. Vis. Image Underst., vol. 63, no. 1, pp. 75–104, Jan. 1996, doi: 10.1006/cviu.1996.0006.

[18] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in In IJCAI81, 1981, pp. 674–679.

[19] C. Zach, T. Pock, and H. Bischof, "A Duality Based Approach for Realtime TV-L 1 Optical Flow," in Pattern Recognition, vol. 4713, F. A. Hamprecht, C. Schnörr, and B. Jähne, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 214–223.

[20] P. Fischer et al., "FlowNet: Learning Optical Flow with Convolutional Networks," ArXiv150406852 Cs, May 2015, Accessed: Jan. 03, 2021. [Online]. Available: http://arxiv.org/abs/1504.06852.

[21] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," Sydney, Australia, Dec. 2013, [Online]. Available: http://hal.inria.fr/hal-00873592.

[22] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow," 2015.

[23] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "DeepMatching: Hierarchical Deformable Dense Matching," ArXiv150607656 Cs, Oct. 2015, Accessed: Jan. 01, 2021. [Online]. Available: http://arxiv.org/abs/1506.07656.

[24] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow," Int. J. Comput. Vis., vol. 92, no. 1, pp. 1–31, Mar. 2011, doi: 10.1007/s11263-010-0390-2.

[25] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A Naturalistic Open Source Movie for Optical Flow Evaluation," in Computer Vision – ECCV 2012, vol. 7577, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 611–625.

[26] T. Pock, M. Urschler, C. Zach, R. Beichel, and H. Bischof, "A Duality Based Algorithm for TV-L 1-Optical-Flow Image Registration," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007, vol. 4792, N. Ayache, S. Ourselin, and A. Maeder, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 511–518.

[27] J. Sánchez Pérez, E. Meinhardt-Llopis, and G. Facciolo, "TV-L1 Optical Flow Estimation," Image Process. Line, vol. 3, pp. 137–150, Jul. 2013, doi: 10.5201/ipol.2013.26.

Verb Sense Disambiguation by Measuring Semantic Relatedness between Verb and Surrounding Terms of Context

Arpita Dutta[1], Samir Kumar Borgohain[2]
Department of Computer Science and Engineering National Institute of Technology Silchar
Assam-788010, India

***Abstract*—Word sense disambiguation (WSD) is considered an AI complete problem which may be defined as the ability to resolve the intended meaning of ambiguous words occurring in a language. Language has complex structure and is highly ambiguous which has deep rooted relations between its different components specifically words, sentences and paragraphs. Incidentally, human beings can easily comprehend and resolve the intended meanings of the ambiguous words. The difficulty arises in building a highly accurate machine translation system or information retrieval system because of ambiguity. A number of algorithms have been devised to solve ambiguity but the success rate of these algorithms are very much limited. Context might have played a decisive role in human judgment while deciphering the meaning of polysemic words. A significant number of psychological models have been proposed to emulate the way the human beings understand the meaning of words, sentences or text depending on the context. The pertinent question that the researchers want to address is how the meanings are represented by human beings in mental memory and whether it is feasible to simulate with a computational model. Latent Semantic Analysis (LSA), a mathematical technique which is effective in representation of meanings in the form of vectors that closely approximates human semantic space. By comparing the vectors in the LSA generated semantic space, the closest neighbours of the word vector can be derived which indirectly provides lot of information about a word. However, LSA does not provide a complete theory of meaning. That is why psychological process modules are combined with LSA to make the theory of meaning concrete. Predication algorithm with LSA was proposed by Kintch, 2001 which was sufficient to capture various word senses and was successful in homonym disambiguation. Meaning of a word might have multiple senses specifically verbs. For example, verb "run" has 42 senses in WordNet. In order to find the correct sense of a verb is really a daunting task and resolving verb ambiguity using psycholinguistic model is very much limited. The proposed method has exploited the high dimensional vector LSA space resulted from training samples by applying predication algorithm to derive the most appropriate semantic neighbours for the target polysemous verb from the semantic space. Finally the vector space of test samples are checked with the training samples i.e. semantic neighbours to classify the senses of polysemous words in accurate manner.**

Keywords*—Word sense disambiguation; ambiguous verb; context; semantic space; latent semantic analysis; polysemy; machine translation

I. Introduction

Even though the research in Word Sense Disambiguation (WSD) has been carried out by researchers from 1940[1] onwards but still the problem is not resolved fully. The ambiguity is present in almost all the natural languages spoken in the world which sometimes makes it difficult to get the correct meaning or sense of a word in the context. Human beings are well organized to understand the meaning of ambiguous words, but in case of machines it requires a mechanism that will help the machine to find out the correct meaning of ambiguous words [2]. For example, ambiguous noun "plane", "The plane flies like a bird in the sky" where the surrounding terms fly, bird, sky can help to recognize the ambiguous term 'plane' is an aeroplane whereas for the example, "the plane is made of paper" where the term paper can identify that "plane" is a geometric plane. Now, if these two examples are given as input text in a computer for machine translation, it is difficult to assume which sense of plane will be considered for translation. If exact meaning of ambiguous term cannot be predicted then the correct meaning of the sentence will be altered. So, the surrounding terms of the ambiguous term must be determined in order to get the true sense of the term. For instance, the word "piggy bank" is related to coin or money that means these terms help to find out the exact sense of bank as it is an ambiguous word.

WSD is one of the most challenging area in the research field of Natural Language Processing dated back to 1940s [3, 4]. There are different approaches to WSD problem such as knowledge base, supervised, unsupervised, semi supervised and hybrid approaches. Knowledge-based approaches were based on different knowledge resources such as machine readable dictionary or thesauri etc. where WordNet [5] is mostly used as a machine readable dictionary in this field. Most of the WSD works are based on different techniques of supervised approach which consists of training and testing dataset. Training dataset of supervised approach is used for classifier to learn and it comprises of target ambiguous words. In contrary, unsupervised approach does not depend on external resources or sense-annotated dataset. Here, word sense discrimination is performed by dividing the occurrences of words into classes to determine the words whether it belongs to the same sense or not. However, Evaluation of unsupervised approach is difficult to measure. Semi supervised approach may be called as minimally supervised approach where unlabelled data is used with the combination

of small quantity of labelled data thereby increasing the machine learning efficiency with much better performance [6]. Hybrid approach is a combination of different types of knowledge resources. In 1950, Kaplan [7] has determined that in a particular context two words on either side of an ambiguous word are equivalent to the whole sentence to the context. Kaplan's work is remarkable in the field of WSD. In 1957, Masterman [8] suggested his theory of finding the actual sense of a word using the headings of the categories present in the Roget's International Thesaurus. In 1964, Yehoshua [9] has pointed out that it is never possible to distinguish ambiguous meaning of a word without a Universal Encyclopaedia as he has used WSD as a part of his machine translation work. In 1980, Searle [10] devised the way in which computer system processed a language. He also highlighted the fact that linguistic symbols are meaningless unless and until it is not grounded or comprehend by someone. In 1990,Miller[11] has invented WordNet which is a revolution in the field of WSD as because there was no such hierarchical organized database of word senses called 'synsets' previously. Later, in 1991, Brown [12] has implemented corpus based WSD for the first time. Needless to say that most of the WSD works are performed on ambiguous noun using different approaches whereas there are very few works available based on ambiguous verb.

One of the significant works on removing ambiguity of verb is predication algorithm [13], which is discussed for homonym disambiguation and similarity judgment by the concept of latent semantic analysis with construction integration model. Another work [14] is highlighted on an interaction between the meaning of a context and vehicle terms of the metaphor where meaning is represented as vectors of a high dimensional semantic space. A new approach [15] has also been discussed on whether multisemantic-role (MSR) based on selectional preferences could be used to improve the performance of supervised verb sense disambiguation method. Here performance is evaluated on two distinct datasets-lexical sample task of SENSEVAL-2 and the verbs from a movie script corpus. Another paper [16] presented one approach to improve the extraction of meaning from Diagnostic corpus by applying little bit modification in predication algorithm. Furthermore, a new concept is proposed [17] where different methods are used to extract the meaning of a polysemic word without using context by vector sum and existing predication algorithm. Some of the distributional approaches that are discussed [18] in literature for sense disambiguation application as well as reformulating the problem of measuring semantic similarity with respect to a particular context and outline a distributional method for identifying diverse documents that activate the sense of polysemous word. One of the reported works [19] has followed the predication algorithm where various semantic space models are compared and also generalized the predication algorithm for the problem of word-concept mapping model from the child learning which is verified by CHILD corpus. In another recent paper [20] on removing ambiguity of noun and verb together where it has discussed the existing methods and also created a new dataset of over 30,000 naturally-occurring non-trivial examples of noun-verb ambiguity for their experiment. This paper also has reported errors that are very often using English part-of-speech tagger related to noun-verb ambiguity. In addition to these, new approach of visual word sense disambiguation for verb senses has been presented in a recent paper [21] by introducing the Multi-Sense dataset of 9,504 images annotated with English, German, and Spanish verbs. They have shown the benefits of cross-lingual verb sense disambiguation model over visual context by comparing uni-modal baselines. In order to find the correct sense of a verb is really a challenging task and resolving verb ambiguity using psycholinguistic model is very much limited. Customarily very few works have been reported on the propose topic but none of them is found to be impressive.

In the proposed work, connectionist network with activation function is used to remove the ambiguity of verb by finding out the surrounding terms to disambiguate the particular sense in which a verb is being used. Here the authors have considered the surrounding words as unordered in nature and found which words are commonly occurred around the target word. This technique is considered as supervised since it requires a training corpus where training must classify each word corresponding to a particular sense. In this work, Latent Semantic Analysis [22, 23] also has been used to find the real meaning of words used in a set of documents as because there is ambiguous term in the document. It maps both the words and the documents into a high dimensional semantic space and finds out the relationship between them. For example, when the word "bat" is used with words like ball, player, field then it may infer a cricket bat. Similarly, the word "bat" with words like trees, wings, fly specifies the sense of "animal bat".

Verb Sense Disambiguation method has not received enough attention in the literature survey of WSD since long time. Most of the WSD work has been performed in different languages using various techniques to remove the ambiguity of noun. There are many databases as well as thesaurus available for noun whereas no proper database is available for verb. Also most of the methods to disambiguate verbs are used in the same way as noun. Therefore, the performance of verb sense disambiguation method is not adequate in the state of art. In this paper, the authors attempt to find the sense of ambiguous verb in a context using vector space model with the notion of activation function to classify senses of verb with most probable surrounding terms despite the lack of conventional verb database.

This paper is organized as follows; Section 2 discusses the methodology of the proposed system. Section 3 discloses the experiment with discussion of various results of the proposed system. Section 4 reveals conclusion and future work.

II. Proposed Methodology

The basic approach of the proposed work is to gather distributional information of high-dimensional vectors and define semantic similarity in terms of vector similarity [24, 25]. Here, authors have used a document as a bag of words (BOW) which is commonly used for information retrieval. In BOW, authors count the number of times each word appears in a document which is the frequency of each word of the document and make a frequency histogram from it. The steps

that are followed by architecture of proposed work can be divided into training phase and a testing phase which is illustrated in Fig. 1.

The schematic diagram in Fig. 1 represents the architecture of work. The proposed methodology may be explained under four broad steps: i) Dataset creation ii) Data pre-processing iii) Training and iv) Testing.

1) Dataset creation: Here authors have perceived the word-sense disambiguation of verbs as a classification task. In any classification task, the machine learning algorithms are applied to a dataset of training samples and later on tested with testing samples. The accuracy of classification depends on the no. of unknown/testing samples that are correctly classified. Since the author's task is classification in nature, a standard dataset of ambiguous verbs only is in demand. However, due to the non-availability of standard dataset of such type; this major challenge is overcome with the creation of custom oriented dataset. Two versions of datasets are created viz. a training dataset containing sentences of ambiguous verbs which are extracted from WordNet and a test dataset which is similar to the former one but the sentences are extracted from Babel Net. WordNet consists of 1, 17,000 synsets/classes organised in the form of a hierarchy. Each of the synset/class has its sense id, gloss and an equivalent example sentence. The information is not limited to the above-mentioned attributes, but carries other information too. But authors have extracted only the sense id, gloss and the example sentences for ten ambiguous verbs namely 'run', 'give', 'break', 'call', 'know', 'put', 'take', 'make', 'draw', 'get'. A total of 500 example sentences encompassing 10 ambiguous verbs are considered in the training dataset. Authors have taken 80% of total dataset as training and 20% as test dataset. Here, example sentences which is depicted as a subset of experimental training dataset for ambiguous verb 'run' are shown in the Table I.

Similar, to the above method of creating the training dataset, authors have created the test dataset from Babel net which contains only the example sentences of the ten ambiguous verbs.

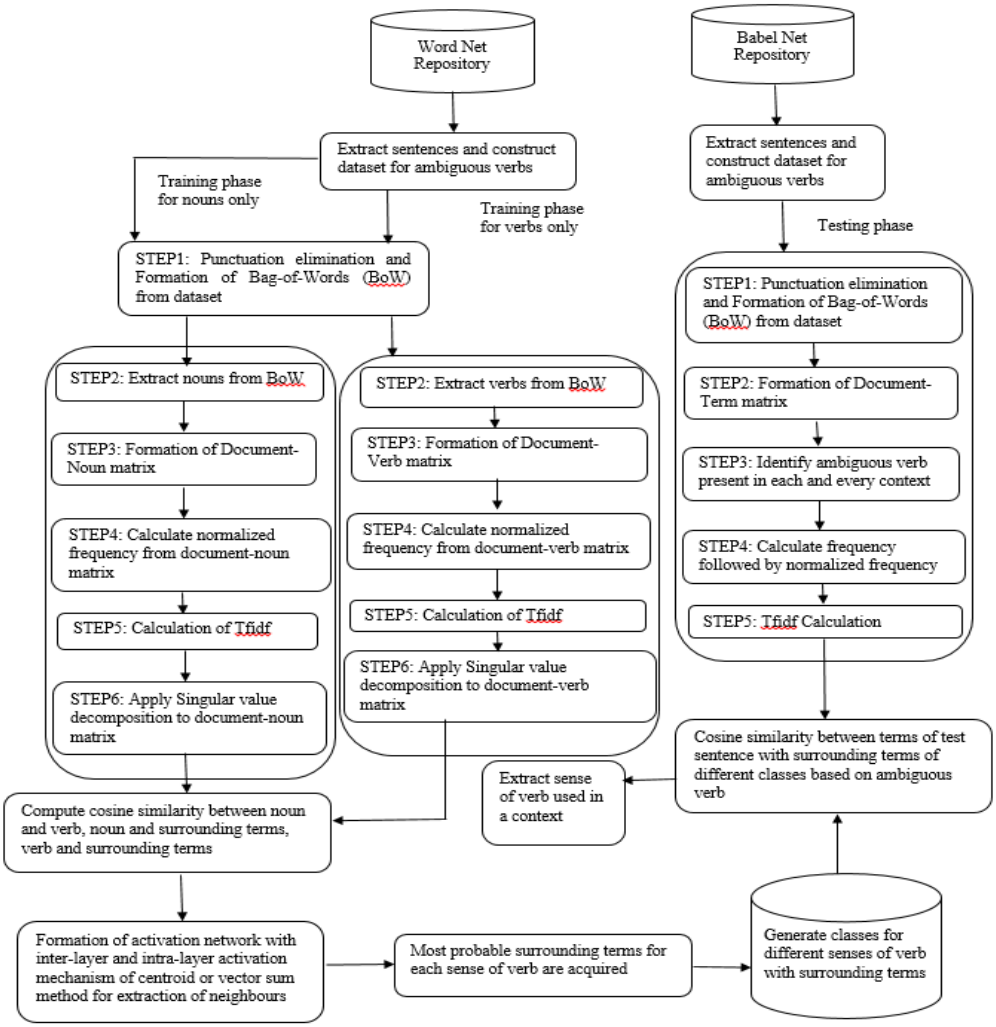


Fig. 1. Block Diagram of Proposed Model.

TABLE I. SUBSET OF TRAINING DATASET FOR AMBIGUOUS VERB 'RUN'

1.	The horse is running in the park.
2.	The horse is running in the race-course.
3.	The horse ran very fast in the race.
4.	The horse runs very fast.
5.	My horse runs last.
6.	The rabbit is running in the garden.
7.	The rabbit runs very fast.
8.	The rabbit is running around in the house.
9.	The kangaroo runs with a baby in its pouch.
10.	The kangaroo is running in the forest.
11.	The machine runs on electricity.
12.	The machine is running in the factory.
13.	The machine runs on crude-oil.
14.	The machine is running very smoothly.
15.	The machine ran properly for many hours.
16.	The colours run.
17.	These dyes and colours are guaranteed not to run.
18.	blood runs in the veins
19.	Blood runs from the heart to all parts of the body through artery.
20.	The bus runs between railway-station and airport.
21.	The bus runs between two states.
22.	The bus runs 100 miles daily.
23.	The bus runs daily throughout the year.
24.	The computer runs the instruction.
25.	The film runs for 3 hours.
26.	The ship runs before the wind.
27.	She runs 10 miles daily.
28.	He runs a new program on the laptop.
29.	They run the tapes over and over again.
30.	He never tires of running that video.
31.	The apple runs large this year.

2) Data pre-processing: As already described in step (i), the attributes contained in the training samples are sense id, gloss and example sentence. In the data pre-processing step, the punctuation markers are removed from the example sentences only. Each of the treated example sentences are then converted to bag-of-words (bow). In case of training phase for nouns, bow contains all unique word tokens including stop words, verbs, pronouns, prepositions, adjectives and numeric values whereas in case for verbs, bow contains all the word tokens. The bow for noun are converted to document-noun matrix and similar to the former approach the bow for verbs are converted to document-verb matrix. Each row of the matrices are considered as a vector having n-dimensions.

3) Training phase: The training phase is divided into two parts, one for training nouns only and another for verb where same steps are followed in both parts. There is a significance of n-dimensional vector representation in the document-verb and document-noun matrix where both the matrices represents this phase involves separate training of n-dimensional noun vector as well as n-dimensional verb vector. In this phase, the tf-idf values of the n-dimensional vectors in the matrices are calculated out. As the terms or words of the dataset are now become vectors, so authors need to compute the weight of all the vectors present in the dataset. As a result of that, term frequency followed by normalized term frequency is computed from which later tf-idf is calculated.

For instance, if the two sentences from training dataset are considered as follows:

S20: The bus runs between railway-station and airport.

S24: The computer runs the instruction.

The bag-of-words formed from the above sentences are:

[*"the", "bus", "runs", "between", "railway-station", "and", "airport", "computer", "instruction"*]

After finding out the term frequency of Bag-of-words, a count matrix is formed where all the sentences of training data set are considered as rows and words as columns shown in Table II.

These frequencies of words of training data set are normalized since each document of training set are of different size. Normalization of frequency is required as because the frequency of a particular word is much higher in a larger document than the smaller document as it contains few terms. Now, the matrices that are built using Latent Semantic Analysis are very large as well as very sparse because most of the cells are blanks due to small number of words in a document. The sparseness of the matrix is removed to get latent features of the terms of Bag-of-words. After that, here documents are converted to vectors of features and finding out the semantic similarity between two documents without considering word order by measuring the distance between these features by cosine similarity. The normalized frequency is obtained by dividing each term frequency with total number of terms present in a document which is shown in Table III.

In reality, certain words that occur too frequently such as article like a, an, the or some prepositions namely of, for, by etc. have little effect in determining the meaning of word. So by weighing down the effects of too frequently occurring words and vice versa for the less frequently occurring words, Inverse Document Frequency is calculated.

TABLE II. COUNT MATRIX FOR PREVIOUS EXAMPLE SENTENCES S20 AND S24

The bus runs between railway-station and airport. The computer runs the instruction.									
S20	1	1	1	1	1	1	1	0	0
S24	2	0	1	0	0	0	0	1	1

TABLE III. NORMALIZED FREQUENCY FOR THE PREVIOUS EXAMPLE SENTENCES S20 AND S24

The bus runs between railway-station and airport. The computer runs the instruction.									
S20	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0
S24	0.4	0	0.2	0	0	0	0	0.2	0.2

IDF= 1 + $\log_e$ (Total number of documents/ Number of documents in which the word is present).

$$TF*IDF=TF*(1+\log_e (N/df)) \quad (1)$$

After training phase of the training dataset taken from Word net ,the words that are very close to the particular sense of ambiguous verb are determined based on their cosine similarity value between nouns and verb. In this way, all the words which are most related to particular sense of a verb are acquired in a class. So, for example, if authors dataset containing ten different senses of an ambiguous verb 'run' then there are ten classes such as 'moving', 'working', 'diffusion', 'flowing' etc. consisting of surrounding words belonging to that class. Authors have chosen sentences for *run* for 10 different senses from Word Net as it is a database which resembles thesaurus. In the similar way, the classes are obtained for other ambiguous verbs present in training dataset. Now, most probable surrounding terms of any particular sense can be increased by adding more sentences in the training dataset. Now, Cosine similarity is calculated between two non-zero vectors after removing sparseness of matrix by singular value decomposition with this formula,

$$similarity = \cos(\theta) = \frac{A.B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

Henceforth, singular value decomposition method is applied because most of the cells are blanks due to small number of words in a document. Cosine similarity is used to find out the similarity between two terms. So, the similarity or distance between ambiguous verbs with surrounding terms, nouns and surrounding terms and ambiguous verb and noun are calculated by cosine similarity. Here, centroid or sum of vector method [15] is used to extract most probable surrounding terms or neighbours. The activation function is tailored by the work of Kintsch [15]. Now, an activation network is formed with three layers where first layer and third layer consist of one node and central layer consists of many nodes. Node in first layer represents ambiguous verb and node in third layer represents noun with which the sense of ambiguous verb will be changed. Nodes in the central layer or middle layer denotes surrounding terms of the contexts which are activated by two activation mechanism such as inter-layer and intra layer activation mechanism [17] .In the inter-layer activation mechanism, nodes in the central layer are activated by both ambiguous verb and noun. In this case, some parameters are used in the formula. Ambiguous verb can be represents as V, noun as N and other surrounding terms in the central layer as O. Therefore, cosine similarity between verb and surrounding terms is Cos (V, O) and cosine similarity between noun and surrounding terms as Cos (N, O).So, activation function for inter-layer of the network would be Inter-layer Activation=Cos (V, O) +Cos (N, O).

In the similar way, Intra-layer activation is calculated where each node in the central layer is inhibited by every one of its neighbours. After computing inter-layer and intra-layer activation of each node in the central layer, an ordering has been done from highest to lowest and first n nodes are chosen accordingly as most probable surrounding terms for the ambiguous verb and particular noun with which verb is used in the context. The activation network is illustrated in Fig. 2.

4) Testing phase: The work has been performed on training data set where sentences consisting of ten different ambiguous verbs such as 'run', 'give', 'break', 'call', 'know', 'put', 'take', 'make', 'draw', 'get' which are taken from Word net. Now, bag-of-words [10] is formed which is a collection of all the words present in the documents. This training data set is prepared with taking care of punctuation and its multiplicity. This bag-of-words concept is used to find term frequency with which a word is appearing in a sentence is considered as a feature point of training. Therefore, a test data set is prepared using Babel net which is a multilingual encyclopaedic dictionary where documents containing ambiguous verbs those are present in training data set. Authors have used Babel Net for formation of testing dataset since it is linked to the computational lexicon of the English language, Word Net. The method which is used to extract feature vectors from training data set, same is used for extracting features from test data set. These features contains nouns and other surrounding terms of the ambiguous verb. To aim is to find the meaning of an ambiguous verb in context as well as the relevant words, semantic similarity of training and test sentences is measured. Whenever, a test sentence comes, the bag-of-words is calculated and ambiguous verb is identified. Therefore, rest terms of the bag-of-words are compared with the terms belong to senses available for the verb. If there is a matching between bag-of-words except verb of test sentence with at least few surrounding terms of same verb of training dataset then cosine similarity score is high.

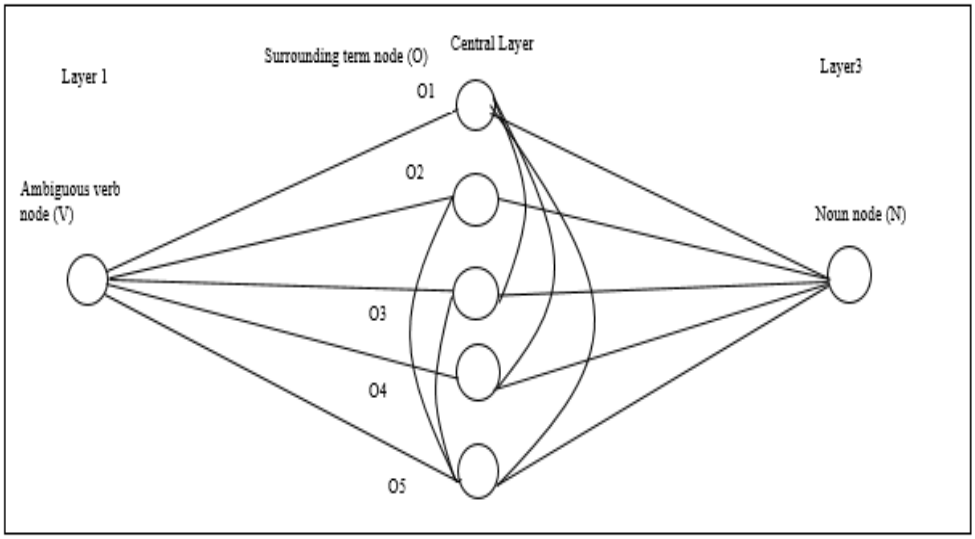


Fig. 2. Layers and Nodes in different Layers of Activation Network.

After that, the sense of the verb can be assumed based on the surrounding terms with which verb is used. Now for example, run is an ambiguous verb as it has multiple meaning depending on the context. It can be used as sense like moving, working, diffusion, flowing, executing, covering a certain distance etc. If in a test sentence containing verb runs with others terms such as machine, electricity, crude-oil which are also present in training document then the result of cosine similarity must be close to 1 indicating the sense for verb 'runs' is working in the test document with those surrounding terms. In the similar way, the terms like blood, artery, heart, veins, body of the test document with verb 'runs' has close to 1 cosine similarly value with the training document indicating that the meaning of runs is diffusion. Now using this method authors can remove the ambiguity of verb in a particular context whenever a new test sentence will come. Ambiguous verb along with the surrounding terms together can solve the problem where the notion of activation function is used in the way that the sense of verb can be achieved if verb is used with those specific surrounding terms of context. The meaning of verb will be completely different depending on the changes of those surrounding terms. In our work, multiple meaning or sense of ambiguous verb as well as training data set has been prepared from Word Net. In the same way, test data set for ambiguous verb is made ready from Babel Net. In our work, both the Training and Test data set are hand-crafted as because there is no such customized verb data set for work till date. So, dictionary cum thesaurus have been used for preparing training and testing dataset.

III. Results and Discussion

The result of the experiment based on the example sentences for ten ambiguous verbs namely 'run', 'give', 'break', 'call', 'know', 'put', 'take', 'make', 'draw', 'get'. Now, for ambiguous verb run can have different meaning such as moving, working, flowing, diffusion, covering a certain distance etc. in different context. Here, authors have shown few example training sentences for run as 'moving' and 'working' sense and also corresponding test context. Now, Table IV shows example sentences for run as moving sense with noun horse and Table V shows example testing sentence. Authors need to find most probable surrounding terms which can appear with run and horse in the same context.

In Table VI, It can be seen that race-course, race, park and last terms have obtained activation value above the range of threshold value 0.5 after applying activation function for the network which consists of verb run, noun horse and surrounding terms like race-course, race, park, fast, last etc. So, it can be concluded that these are the most probable surrounding terms in a context if ambiguous verb run is used with noun horse. These surrounding terms can be increased if more number of sentences are trained which is containing verb run with noun horse with it. At the same time, run and horse together can infer the sense of run is moving provided race or race-course or last appears together in the same context.

If any test context or query sentence which is containing verb run with horse, then similarity between surrounding terms of test context with any of the surrounding terms which have obtained in Table VI can help to find the sense of ambiguous verb run. So, different classes with surrounding terms can be formed after training the dataset containing verb run. Authors can also find cosine similarity between query document with any document of training dataset containing same verb and noun which is shown in the above Table VII. If cosine similarity value is close to 1 that indicates the surrounding terms of test context is similar with training document. Table VIII and Table IX display example training sentences and testing sentence for 'run' as moving sense, respectively.

TABLE IV. Example Training Sentences for Run as 'Moving' Sense

1. The horse is running in the park.
2. The horse is running in the race-course.
3. The horse ran very fast in the race.
4. The horse runs very fast.
5. My horse runs last.

TABLE V. EXAMPLE TEST SENTENCE FOR RUN AS 'MOVING' SENSE

1. The horse runs faster than zebra in race. The horse runs the race for the cup.

TABLE VI. ACTIVATION VALUE TO FIND SURROUNDING TERMS FOR MOVING SENSE OF VERB RUN

neighbours of *run* for 'moving' sense	**Cosine neighbour of *run***	**cosine neighbour of *horse***	**activation value**
race-course	0.350785733	0.358330308	0.709116041
race	0.658504608	0.31353902	0.972043628
park	0.24804297	0.253377791	0.501420761
fast	0.266184711	0.126740637	0.392925348
last	0.24804297	0.450906457	0.697721584
very	0.148172217	0.085139467	0.233311684

TABLE VII. COSINE SIMILARITY TO FIND SENSE OF AMBIGUOUS VERB IN TEST SENTENCE

Neighbourhood of ***run*** in **training dataset of Table IV**	Neighbourhood of ***run*** in **test sentence of Table V**	Semantic similarity between training sentences and test sentence 1
horse	horse	0.9800
race	race	
fast	faster	
last	zebra	
park	cup	
race-course		

TABLE VIII. EXAMPLE TRAINING SENTENCES FOR RUN AS 'WORKING' SENSE

11. The machine runs on electricity. *12. The machine is running in the factory.* *13. The machine runs on crude-oil.* *14. The machine is running very smoothly.* *15. The machine ran properly for many hours.*

TABLE IX. EXAMPLE TEST SENTENCE FOR RUN AS 'WORKING' SENSE

1. The machine runs on electricity in the factory for whole day.

In Table X, It can be seen that crude-oil, electricity, factory etc. terms have obtained activation value above the range of threshold value 0.5 after applying activation function for the network which consists of verb run, noun machine and surrounding terms like crude-oil, electricity ,hours, many, factory, properly etc. Now, run and machine together can infer the sense of run is working provided factory, crude-oil, electricity etc. appear together in the same context. Table XI shows the cosine similarity between training and testing sentence for verb 'run' as working sense.

Table XII, XIII and XIV show example training sentences, testing sentence and activation value to find surrounding terms respectively for verb 'run' as covering a certain distance. If any test context or query sentence which is containing verb run with machine, then similarity between surrounding terms of test context with any of the surrounding terms which we have obtained in Table X can help to find the sense of ambiguous verb run. Confusion matrix is used to measure the performance of machine learning classification which is around 0.8235. It can be improved by increasing number of sentences in the dataset for ambiguous verb with all available sentences.

TABLE X. ACTIVATION VALUE TO FIND SURROUNDING TERMS FOR WORKING SENSE OF VERB RUN

neighbours of *run* for 'working' sense	**Cosine neighbour of *run***	**cosine neighbour of *machine***	**activation value**
crude-oil	0.274597701	0.51883493	0.793432631
electricity	0.274597701	0.51883493	0.793432631
smoothly	0.409250021	0.432362441	0.841612463
factory	0.350785733	0.370596378	0.721382111
properly	0.752576695	0.370596378	1.123173073
hours	0.145619993	0.202143479	0.347763472
many	0.398588494	0.1962796	0.594868094
very	0.148172217	0.156540252	0.304712469

TABLE XI. COSINE SIMILARITY TO FIND SENSE OF AMBIGUOUS VERB IN TEST SENTENCE

Neighbourhood of ***run*** in **training dataset of Table VIII**	Neighbourhood of ***run*** in **test sentence of Table IX**	Semantic similarity between training sentences of Table VIII and test sentence 2	Semantic Similarity between training sentences of Table IV and test sentence 2
machine	machine	1.0000	0
smoothly	electricity		
electricity	factory		
factory	whole		
properly	day		
hours			
crude-oil			

TABLE XII. EXAMPLE TRAINING SENTENCES FOR RUN AS 'COVERING A CERTAIN DISTANCE' SENSE

20. The bus runs between railway-station and airport. *21. The bus runs between two states.* *22. The bus runs 100 miles daily.* *23. The bus runs daily throughout the year.*

TABLE XIII. EXAMPLE TEST SENTENCE FOR RUN AS 'COVERING A CERTAIN DISTANCE' SENSE:

3. The bus ran 10 miles that day.

TABLE XIV. ACTIVATION VALUE TO FIND SURROUNDING TERMS FOR COVERING A CERTAIN DISTANCE SENSE OF VERB RUN

neighbours of *run* for '*covering a certain distance*' sense	cosine neighbour of *run*	Cosine neighbour of *bus*	activation value
miles	0.2745	0.4511	.7256
daily	0.1140	0.5941	.7081
states	0.2288	0.4511	.6799
railway-station	0.1716	0.3866	.5582
airport	0.1716	0.3866	.5582
year	0.1548	0.2245	.3793

IV. CONCLUSION AND FUTURE WORK

Removal of ambiguity of polysemous verb is very hard as it depends on the context. If the context of the same verb is altered then the meaning of the verb will be different. Since, ambiguity of verb needs to be removed in machine translation as inappropriate translation of source always leads to misprediction of information. In this work authors have used supervised machine learning approach by using centroid or vector sum method which helps in finding the meaning of ambiguous verb by classifying different senses of an ambiguous verb with most probable surrounding terms with it. So an ambiguous verb can be used with particular words for the specific sense of that verb and surrounding terms are changed if the sense of that verb in context is different. Therefore, sense of the verb can be predicted based on most probable surrounding terms only. Efficiency of the proposed method can be increased with larger set of data as more surrounding terms can be obtained which are feature points of the ambiguous verbs. The authors have used combination of dictionary and thesaurus for acquiring senses available for ambiguous verb as well as preparing hand crafted dataset for training and testing since there is no dataset available for verb. Future work may include increasing the accuracy of this method with the creation of database for ambiguous verb with all available senses.

ACKNOWLEDGMENTS

The authors are highly thankful to the Ministry of Electronics and Information Technology, Government of India initiated "Visvesvaraya PhD Scheme for Electronics and IT" for the support to carry out the research work.

REFERENCES

[1] Daniel Jurafsky,James H. Martin., "Speech and language processing: an introduction to natural language processing, computationallinguistics, and speech recognition",2nd ed. Pearson Prentice Hall, Inc., Upper Saddle River, NJ, 2009.

[2] A.R.Pal, D.Saha, "Word sense disambiguation: A survey," International Journal of Control Theory and Computer Modeling, vol.5, no.3, July 2015.

[3] Word Sense Disambiguation; Algorithms and Applications, Edited by Eneko Agirre and Philip Edmonds, Springer, VOLUME 33.

[4] Zipf, George Kingsley. "Relative frequency and dynamic equilibrium in phonology and morphology", In Proceedings of the 6th international congress of linguists. Paris, 391-408, 1949.

[5] George A. Miller. Word net: A Lexical Database for English Communication, ACM, 38(11):39-41, 1995.

[6] Nandanwar, Lokesh, and Kalyani Mamulkar. "Supervised, semi-supervised and unsupervised WSD approaches: An overview." International Journal of Science and Research (IJSR) 4.2 (2015): 1684-1688.

[7] A. Kaplan. An experiment study of ambiguity and context. Mechanical Translation, 2:39–46, 1955.

[8] Masterman, Margaret. "The thesaurus in syntax an semantics." Mechanical Translation 4.1-2 (1957): 35-43.

[9] Y. Bar-Hillel. The present status of automatic translation of languages. In Advances in computers, volume 1, pages 91–163. Elsevier, 1960.

[10] J. R. Searle. Minds, brains, and programs. Behavioral and brain sciences, 3(3):417–424, 1980.

[11] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to word net: An on-line lexical database. International journal of lexicography, 3(4):235–244, 1990.

[12] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. Word-sense disambiguation using statistical methods. In Proceedings of the 29th annual meeting on Association for Computational Linguistics, pages 264–270. Association for Computational Linguistics, 1991.

[13] Kintsch, W. (2001)," Predication", Cognitive Science, 25, 173-202.

[14] Kintsch, Walter. "Metaphor comprehension: A computational theory." Psychonomic bulletin & review 7.2 (2000): 257-266.

[15] Ye, Patrick, and Timothy Baldwin. "Verb sense disambiguation using selectional preferences extracted with a state-of-the-art semantic role labeler." Proceedings of the Australasian Language Technology Workshop 2006. 2006.

[16] Jorge-Botana, Guillermo, Ricardo Olmos, and José Antonio León. "Using latent semantic analysis and the predication algorithm to improve extraction of meanings from a diagnostic corpus." The Spanish journal of psychology 12.2 (2009): 424-440.

[17] Jorge-Botana, Guillermo, et al. "Visualizing polysemy using LSA and the predication algorithm." Journal of the American Society for Information Science and Technology 61.8 (2010): 1706-1724.

[18] Rumshisky, Anna. "Resolving polysemy in verbs: Contextualized distributional approach to argument semantics." Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics/Rivista di Linguistica (2008).

[19] Kievit-Kylar, Brent, George Kachergis, and Michael Jones. "Naturalistic word-concept pair learning with semantic spaces." Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 35. No. 35. 2013.

[20] Elkahky, Ali, et al. "A Challenge Set and Methods for Noun-Verb Ambiguity." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[21] Gella, Spandana, Desmond Elliott, and Frank Keller. "Cross-lingual Visual Verb Sense Disambiguation." arXiv preprint arXiv: 1904.05092 (2019).

[22] Dumais, Susan T., et al. "Using latent semantic analysis to improve access to textual information." Proceedings of the SIGCHI conference on Human factors in computing systems. 1988.

[23] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25.2-3 (1998): 259-284.

[24] Kintsch, W. (2007) Meaning in context, in T. K. Landauer, D. McNamara, S. Dennis and W. Kintsch (eds) Handbook of latent semantic analysis. Mahwah, NJ: Erlbaum.

[25] Soumya George K, Shibily Joseph, Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. V (Jan. 2014), PP 34-38.

Water Level Monitoring and Control System in Elevated Tanks to Prevent Water Leaks

Christian Baldeon-Perez[1]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Perú

Brian Meneses-Claudio[2]
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades
Lima, Perú

Alexi Delgado[3]
Interdisciplinary Research Center Science and Society (CIICS)
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—The shortage of water is recurrent in Lima, Perú and in the world, either due to natural disasters, deficiency in pipes due to age, or breakages by external agents such as heavy trucks, heavy machinery, etc., which damage to the underground pipes causing flooding and, shortages in the affected area or zone. As a possible solution, many inhabitants have elevated tanks, but they do not have an automatic control, nor view the water level in the tank, nor recognize possible water leaks if they occur, such leaks are economic detriment to the user. The objective research work wants to avoid shortages in the short period, controlling and monitoring of water for use at home or industry. For the implementation of this project, the technology of the Arduino Uno board, a 16x2 LCD screen, an ultrasonic sensor, and a mini pump will be used, which will be fed with a DC voltage, it is intended to have manual control every time, otherwise the work will be fully automatic. The results obtained were as expected, always displaying on the LCD screen, as at the beginning of the process with the tank empty and its corresponding alarm on a led diode, the percentage of water that gradually rises until the end of the process with the full tank message and its corresponding alarm on another led diode. The implementation of this project is economical, which is why it is very viable for many households and companies that can choose this alternative.**

***Keywords*—Arduino; ultrasonic sensor; stockouts; algorithm; monitoring and control; water leaks; elevated tanks**

I. Introduction

Many homes, residential, condominiums, etc., have water reservoirs on the rooftops, but they do not have a constant monitoring system that guarantees optimal performance to visualize the status of storage or water leakage problems which could arise due to natural disasters [1], or due pipe breaks caused by heavy means of transport or due the age of the canalization system [2].

The fluidity of water is essential to carry out any activity or work, but since this objective is achieved in times of shortage or problems in the supply of drinking water, it is convenient to make an underground well and elevated tanks to counteract this adversity [3].

There are companies that depend on great fluidity of water for their different processes, such as laundries or dry cleaners, how can the constant supply of the liquid element be guaranteed, in addition to saving on the implementation of a constant control and monitoring of tanks elevated, powered by constant pressure pumps and PID control [4].

The main objective of this project is to avoid water shortages and leaks either in houses, residences, condominiums, and companies whose area is causally linked to the use of water such as the different textile companies, laundries, and dry cleaners. With this project, the use of water will be efficiently improved for its different needs and the low budget it manages, since the electronic elements to be used are low cost and very accessible in the market, also mention that this project will prevent water leaks [7].

To develop this research work, we will use the open-source electronic creation platform (Arduino), which is based on free hardware and software, flexible and quite easy to use to create and develop. To create the pseudocode, we will go to the Arduino IDE programming language, which is based on the C ++ language. It will also need an element that is capable of measuring distance to objects, such as the ultrasonic sensor, which in this case will measure distance to water and will be processed on the Arduino board and displayed on the LCD screen.

The following research work is structured as follows: In Section II, the electronic elements to be used will be shown such as the Arduino board, ultrasonic sensor, LCD screen, among others, in addition to the pseudo-code that will make possible the operation of the circuit to be put testing. In Section III, the results obtained, the complete layout, step by step of the water percentages and the corresponding alarms such as: empty tank and full tank will be shown, which will verify the good design and application of the project. In Section IV, the discussions regarding similar works will be shown, but that in comparison to the presented project differ in the number of accessories and operating costs. In Section V, the respective conclusions and recommendations of the future work that will be achieved with this work will be appreciated.

II. Literature Review

Using free software and hardware, the process of monitoring water level and leaks can be efficiently automated as presented by Astudillo [5], an automated system with application of an ultrasonic sensor, LCD screen and Arduino Mega board as the main elements for the control of the aforementioned process, as a result an efficient monitoring

system was obtained in a condominium in which it was properly implemented, the water level is displayed at all times on the LCD screen and if it exceeds the higher level, a water leak alarm is displayed.

Also using a low budget and guaranteeing optimum performance, the Arduino board with an ultrasonic sensor and solar powered can be used to avoid the need to resort to the 220v alternating current, as presented by Núñez and Martínez in a house in Lorica, Córdoba in Colombia, achieving successful independence of the entire project from the alternate mains supply, also achieving real-time monitoring with the visualization on the LCD screen of the percentage of water in the elevated tank and the alarms programmed into the Arduino Mega board. It is an excellent alternative to monitor in real time and specify the status of the water level in elevated tanks [6].

In the project presented by Valderrama [4], where alternate water pumps are connected to supply, in addition to adding a Siemens PLC to work directly with the touch screen, solenoid valves for better handling of all processes, this project was implemented for a thesis at the Universidad Católica de Colombia, achieving the total control of the processes through the touch screen.

III. Methodology

The type of research is Experimental, it seeks to apply scientific theoretical knowledge to the solution of a practical and immediate problem of knowledge through the implementation, transformation and/or modification of concrete reality [8]. In this sense: The methods that we will use during the research process are Deductive and Scientific. Deductive Method: It is the one we use to explain the characteristics of the technology with Scratch architecture and electronic prototype platform - ARDUINO. Scientific Method: We will use this method to define our concepts, hypotheses and variables that gives us the resources and intellectual instruments to solve the problem [9].

As has already been reflected previously, the basic objective of this project is to provide users with real information on the status of the elevated tank in the place of residence, this task will be broken down from three different computers: The first, the Arduino device installed as the operations center of the water tank and that will serve as receiver and transmitter of all information. The second, the electronic ultrasound device installed on the top of the tank, which will always inform the Arduino system of the status of the water level. The third, an LCD display to view each information processed by the Arduino [10].

A. Software

For the project, it will use the Scratch program. Which is very compatible with Arduino and it will give below some scopes of it: Scratch 4 Arduino, S4A, developed by the Spanish Citilab. It has in its favor that for practical purposes it is a Scratch 1.4 modified to allow connection with an Arduino board and add the corresponding blocks that allow interaction with it. Developed and published since 2010, it is not updated as frequently as would be desirable, but despite this, it allows to work with standard Arduino boards without complications and in a familiar environment, such as the original Scratch 1.4. Completely free and available for Windows, macOS and Linux systems [11].

1) Arduino pseudocode: The model of the board to use will be the Arduino Uno, which is an electronic board based on the ATmega328 microcontroller. It has 14 digital inputs/outputs, of which 6 can be used as PWM (Pulse Width Modulation) outputs and another 6 are analog inputs [6].

a) Valve opening and closing: The minimum water percentage will be 10% and the maximum will be 90%. These values will be very important since they will start and end the filling and monitoring of the water tank [12]. The process will begin with the start-up through the mini start button, at that moment the mini water pump will begin to work, making it possible to fill and register the water level through the ultrasonic sensor, which will be important since it measures the distance to the liquid, using ultrasonic waves, the head emits an ultrasonic wave and receives the reflected wave that returns from the water. Ultrasonic sensors measure the distance to the object by counting the time between emission and reception. and displayed through the LCD screen, which will culminate when reaching 90% at which time the work of the water pump will be cut off, it should be noted that if it exceeds 90% the system will send an alarm to check the system [12].

b) Verification and monitoring of water level: Water level monitoring will be made possible by the ultrasonic sensor and 16x2 LCD screen (the term 16x2 LCD refers to a small device with a liquid crystal display that has two rows, of sixteen characters each, used to display information, usually alphanumeric), which will show in real time the percentage of water in the tank [3].

c) Water flow abnormality alarm: The leak alarm can be displayed on the LCD screen when the percentage of water level exceeds 90%, at that moment through the programming entered the Arduino it will make it possible to send a voltage to an optocoupler integrated circuit to activate the lamp. circuit failure, for which an alarm will be issued to show the system failure in order to fix it [13].

B. Hardware

1) System Implementation: The model will consist of a plywood base, in which it will be conditioned in a cylindrical container that will act as a water tank, a 12V DC mini water pump, which will allow the filling of the tank, an ultrasonic sensor, which will be located in the upper part of the water tank and will allow data to be sent to the Arduino board of the status of the water level, an LCD screen, in which the percentage of water can be read, 1 mini start button , which will allow the start of the process and 1 mini stop button, to stop the action at any point in the process [14].

The power supply will be important for the good performance of the components, in Fig. 1 we can clearly see the connection between them. The use of the ultrasonic sensor is based on the emission of an ultrasonic pulse to a reflective surface, the free surface, and the reception of its echo in the

receiver [15]. Also, in Fig. 2, there is the connection of the mini pump to the activation circuit that is why the present project is automatized so when the system requires water, it will activate the mini pump.

The trigger circuit will be very important because it will govern the pump when it needs to be enabled with the Arduino [16].

Process: As shown in Fig. 3, the implementation will be made in a 30cm x 50cm plywood model, previously conditioned, then the mini water tank with a built-in water vent will be fixed, then a protoboard will be fixed to put the LCD screen and the Arduino board to later wire them, LED diodes will also be fitted, which will indicate processes such as: process running, low level, medium level, full tank.

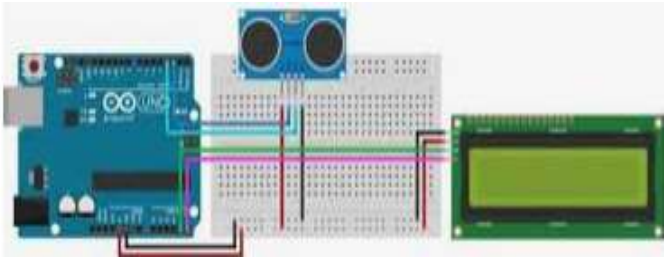


Fig. 1. Communication between Arduino, Ultrasonic Sensor and LCD Screen.

Fig. 2. Connection of the Mini Pump to the Activation Circuit.

Fig. 3. Breadboard with Arduino and LCD Screen Wired

Fig. 4. Water Container.

As shown in Fig. 4, the manufactured container will act as a water tank in which the liquid will be stored and then discharged through the discharge valve.

IV. Results

A. Program Algorithm

To program the microcontroller, the Arduino IDE is used, which will make it possible to govern and activate what is necessary for the system to operate well at each stage of the process [12].

As we appreciate in part of the program below, the activation and deactivation of the mini pump is clearly appreciated, which will be left indicated to be able to start the process.

```
#include<liquidCrystal.h>
Cont int rs=13, en=12, d4014, d5=27, d6=26, d7=25;
LiquidCrystal lcd(rs, en, d4, d5, d6, d7);
Byte bombaOn[8]={
0b00100,
0b01110,
0b11110,
0b00100,
0b11111
};
Byte bombaOff[8]={
0b11111,
0b00100,
0b11110,
0b01110,
0b00100
```

In the part of the program below, we proceed to give some details of the program for the work of the ultrasonic sensor, which is programmed the range in cm and the percentage is assigned, it also shows which output pins it will work on [12].

```
#define DIST_TOPE120// maximum level, measured with empty tank

const int trigPin =2;
const int echoPin =5;
float distancia = 0;
int nivel = 0; // level of percent
```

In part of the program below, the assignment of percentages to the different levels of flow perceived by the ultrasonic sensor and sent to the Arduino Uno board is shown. The activation of the relay on the Arduino board to cut the water pump in the moment that the maximum water storage level is reached [16].

```
Int ultrasonic_fail = 0
Void get_leve( ){
Distance = get_dist();
If (distance == 0) ultrasonic_fail++;
If (distance_fail ==5){
DigitalWrite (CTROL_RELAY_GPIO,LOW);
Lcd.clear ();
Lcd.print ("TANK FULL");
While (1){
}
If (distance > 0){
```

B. System Implementation

The tank monitoring system, in percentages of water and detection of failure in case of water leakage is displayed in Fig. 8. As the reader can see, in Fig. 5, there is the prototype of water monitoring system, also the connections of the Arduino board, display and the ultrasonic module.

Initially, 7V DC is supplied for the correct operation of the different stages of the system. At that moment, the message "EMPTY TANK" will be displayed on the LCD screen, which will start the mini water pump, thus starting the filling of the tank, as shown in Fig. 6.

At the beginning of the process, an ultra-bright LED diode was also incorporated as an indicator in case the LCD screen could suffer from any damage and no message could be displayed as seen in Fig. 7.

Following the process, the percentage is displayed in real time at each instant of the tank filling as seen on the LCD screen, additionally an ultra-bright LED was incorporated, which will indicate that the filling is being carried out in case of failure in the tank. LCD screen and could not be viewed on the screen, as shown in Fig. 8, 9 and 10.

As can be seen in Fig. 10, the process is carried out normally, clearly indicating the percentage of water.

As we can see in the previous figures, the work of the water percentage sensor is what is expected, finally when it passes 90% on the LCD screen the message “TANK FULL” will be displayed as the reader can see in Fig. 10, at the same time it will turn on an amber led, time in which the filling process will be concluded, at this moment the water pump will stop working until the water level drops to 20%, with which the process will start again.

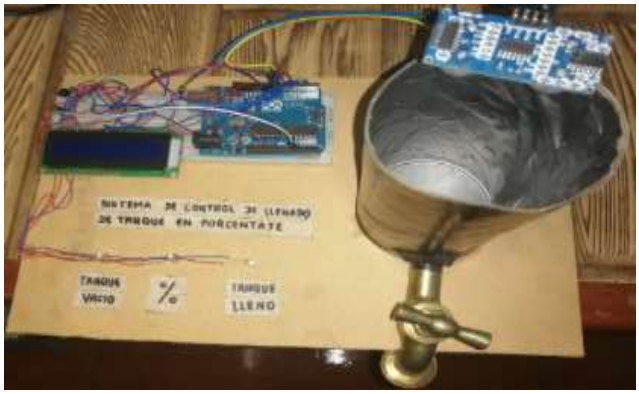


Fig. 5. Water Monitoring System

Fig. 6. Starting the Process

Fig. 7. Process Start Indicator Led.

Fig. 8. Tank Fill to 20%.

Fig. 9. Tank Fill to 60%.

Fig. 10. Tank Full.

As can be seen in each image, the work done by the prototype implemented to control the water level and failures in elevated tanks is as expected. The full-scale implementation is very economical, an economic estimate of 200 Peruvian Soles, for an automatic and modern system for efficient work, which each household should have implemented (economic, modern and safe).

The prototype of water level monitoring was tested in an electronic laboratory from Universidad de Ciencias y Humaniades, proving their running in differents floor of the university. As a future work, we want to implement this research work in different house and heights.

V. Discussions

The Project that is presented in this Paper has specific and necessary functions that are required for the proper functioning together of how to monitor the elevated water tank and its possible breakdowns, which is not the case with the project provided with Morales and Flores [8], in which the program is very extensive to describe each working moment, in addition to adding several loops to complete the functionality of accessories such as solenoids, solenoid valves, limit switches, pumps and control relays, which are many . Comparisons are to scale in working mode of the accessories used.

In terms of programming, we also see substantial differences between the work presented by Morales, Flavio and Flores [8], they created a web page to see in real time on the pc and on their cell phones, the operation of their project, using the following languages programming: HTML, PFP, JAVASCRIPT and CSS, on the other hand, the program presented in this Paper only has a C ++ type language, in addition to showing, with simple steps, how a good performance of the circuit is achieved at the scale of monitoring elevated tanks.

The project presented by Navarro [16], is very interesting, in which a work like the one presented in this research work is described on a real scale, it shows us the programming in blocks (Mblog), which is easily understood with a program since it is short and simple (a single loop), in terms of materials it uses almost the same as those described in this research work (ultrasonic sensor, LCD screen, etc.), also it uses an electronic card (D1 Mini Chipset), which calculates the distance and connects via Wi-Fi, it also has power via a solar panel and the voltage is stored in lithium batteries, which makes this project independent from the home alternating power network.

If this research work would need to be implemented in a building with 15 floors or more, it would be necessary to incorporate components and accessories similar to that presented by Astudillo [5], in which a more extensive circuitry is appreciated, adding limit switches, solenoid valves, PID control and a complex algorithm with 5 loops, ideal for jobs where more control and electronic security are required for the inhabitants, compared to our research work that is implemented for a 3-story house.

Castillo [17] shows in his work where, in addition to the accessories used in the present research work, a touch screen is displayed, which makes the entire process very manageable, compared to ours, since we only use the LCD screen, in which nothing can be accessed since it is only visualization, from the touch screen you can enter everything necessary for the operator to control all the functions that needed, a great achievement of links and algorithms. It should be noted that, in economic terms, this system is expensive since the work presented in this research work would be only 15% of the cost of the Castillo project, of which, budget increase due to the use of the touch screen, there will always be people or companies with the economic budget to acquire them.

We also have the option of Valencia [18], who uses in his project the SIEMENS NANO PLC LOGO, which uses more electromechanical components such as buoys, more voltage power supplies, more relays to enable and disable the process, more lamps of Signaling, compared to our research work, which only uses the Arduino since it is very programmable and for what is required in this project is more than enough and the same results obtained by Astudillo are obtained, it is undeniable that by reducing the circuit part to electronic mode, it is more economical and practical to make this project more viable.

VI. Conclusions

The results obtained were as expected, from the beginning the problem was identified, in this case foreseeing water shortages either naturally, caused by man or deterioration of the pipes over time, the use of elevated tanks, is beneficial for these cases mentioned above, we add constant monitoring and warning in case of leaks or water shortages, also the project is very viable.

The process begins with the tank empty, at that moment the system will launch the alarm and the mini water pump will start to work, the percentage of water will rise and be displayed on the LCD screen, through the ultrasonic sensor it will be possible this job of sending information to the Arduino board and sent for viewing on the LCD screen mentioned above, the percentage of water will rise until it reaches 90% of the water level, after that the process will stop until the level drops to 20%, instant in which the process will resume again, it should be noted that if the water level exceeds 90%, the system will emit an alarm with the word "TANK FULL" which will warn of a possible leak so, unnecessary expenses in monthly billing will be avoided.

The system shown, designed and tested guarantees optimal performance, the theory learned in university classes and subsequent research are put into practice, which are essential for the development of this project, it should be noted that a real problem arises and its possible solution for certain days of water shortages, it is also mentioned how economical this system is for the home and that depending on the area to work, more elements can be added to improve its performance, such as: solar power sources , PLC, touch screen, among others.

As a future work, it will be sought to disseminate it and manufacture it to be able to commercialize it, since, as can be seen in this paper, it is ideal, economical, and exact in monitoring the water level in elevated tanks for standard homes and with some more additional accessories, it will also be possible to use it in companies.

REFERENCES

[1] Consorcio Evaluación de Riesgos Naturales – América Latina, Perú: Gestión de Riesgo de Desastres en Empresas de Agua y Saneamiento, 1st ed., vol. 1. Lima - Perú: Programa de Agua y Saneamiento del Banco Mundial, 2012.

[2] M. Medina, "Sedapal pierde 28% de su facturación por tuberías rotas y redes clandestinas," Correo, NOTICIAS CORREO, Lima - Perú, pp. 1–2, Jun. 06, 2017.

[3] J. Perez, "Automatic Supply, Tanking and Leak Detection System in a Long Distance Drinking Water Network using ZIGBEE Technology," Universidad Autonoma del Estado de México, Atlacomulco - México, 2019.

[4] K. Bohorquez, D. Fonseca, and S. Gutiérrez, "Didactic System for Level Control with Coupled Tanks," Universidad Católica de Colombia, Bogotá - Colombia, 2017.

[5] R. Astudillo, "Design and Implementation of a Prototype of a Water Level Meter through an Ultrasonic Sensor for Depressed Steps," Universidad Tecnológica Israel, Quito - Ecuador, 2016.

[6] Y. Núñez Tordecilla and M. Martínez Vélez, "Design and implementation of a system for the control of water consumption based on Arduino technology and controlled through a web application in a house in the municipality of Lorica Cordoba," Facultad de Ingeniería, Córdoba - Colombia, 2018.

[7] "Exportando con Expoberto," Lima-Peru, 2018.

[8] Morales Arévalo. Flavio and L. Flores Valencia, "Implementation and Monitoring of an Automatic Control System by means of Arduino GPRS board for a Hydraulic Pumping System installed in a well and in two drinking water pumping stations," Quito, Quito - Ecuador, 2019.

[9] W. U. A. Depaz Sandon, "Prototype using Arduino technology for water level measurement in dangerous lagoons of the Huascarán National Park. 2018," Universidad Nacional Santiago Antúnez de Mayolo, Huaráz - Perú, 2018.

[10] A. Ríos Caicho and R. Peñafiel Adrián, "Prototype of an automated system for small-scale industrial applications based on a mobile graphical interface controlled by arduino," Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería En Networking y Telecomunicaciones, Guayaquil - Uruguay, 2018.

[11] Aprendiendo Arduino, "Lenguaje de programación C++," Wordpress, 2018. https://aprendiendoarduino.wordpress.com/2015/03/26/lenguaje-de-programacion-c/ (accessed Feb. 23, 2021).

[12] A. Mejía Palomino, "Design and simulation of a dehumidification system for diesel bulk storage tanks controlled by means of an Arduino board," GUAYAQUIL/UIDE/2018, Guayaquil - Ecuador, 2018.

[13] Arduino Official Store, "Arduino Uno Rev3," Arduino Official Store, Mar. 02, 2016. https://store.arduino.cc/usa/arduino-uno-rev3 (accessed Feb. 23, 2021).

[14] W. Alvarado-Diaz, B. Meneses-Claudio, V. Romero-Alva, O. Minaya-Varas, A. Roman-Gonzalez, and M. Zimic, "Design of a system of analysis of bioimpedance in children," Jan. 2019, doi: 10.1109/ICA-ACCA.2018.8609844.

[15] B. Meneses-Claudio, W. Alvarado-Díaz, F. Flores-Medina, N. I. Vargas-Cuentas, and A. Roman-Gonzalez, "Detection of suspicious of diabetic feet using thermal image," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 379–383, 2019, doi: 10.14569/IJACSA.2019.0100648.

[16] J. Navarro, "Remote water level monitoring in a home tank using ultrasound," Medium, Aug. 26, 2018. https://medium.com/@jnvrro/monitoreo-remoto-de-nivel-de-agua-en-tanque-hogareño-usando-ultrasonidos-2384857906dd (accessed Feb. 23, 2021).

[17] J. Castillo Ruiz, "Design and Electronic Development of the Automatic Control of Filling and Emptying of a Rapid Compression machine," Universitat Politècnica de València, Valencia - España, 2017.

[18] C. Valencia Aguilar, "Design of a Level Monitoring System for EMCALI Telecomunicaciones Emergency Tanks," Universidad Autónoma de Occidente, Santiago de Cali - Colombia, 2013.

An Evaluation of the Localization Quality of the Arabic Versions of Learning Management Systems

Abdulfattah Omar
Department of English
College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University, Egypt

Abstract*—The recent years have witnessed the development of numerous Learning Management Systems (LMSs) to address the increasing needs of individuals and institutions all over the world. For accessibility and commercial purposes, many of these LMSs are released in different languages using what is known as localization systems. In this regard, there has been a parallel between the development of LMSs on one hand and the localization systems on the other. One main aspect in the recent evaluation systems and studies of LMSs is localization quality. Despite the prolific literature on localization quality, very little has been done on Arabic localization. As thus, this study is concerned with the evaluation of the localization quality of the Arabic version of LMSs. In order to explore users' perceptions towards the Arabic versions of the LMSs, an online questionnaire was conducted. Participants were asked about their familiarity with LMSs and whether they used the Arabic versions of these systems. They were also asked about their experiences with the Arabic localization of these systems and whether they faced any problems in dealing with the Arabic version. The findings indicate that translation inconsistencies are the main problems with the Arabic versions of different LMSs including Blackboard Learn, Microsoft Teams, and Zoom. These problems have negative impacts on the effectiveness and reliability of these systems in schools, universities, and training institutions. For the proper implementation of LMSs both localization and translation should go hand-in-hand. Localization developers and LMSs designers need to consider the linguistic the peculiar linguistic features of Arabic. The findings of the study have implications to translation programs in Arab universities and training institutions. Program designers should integrate translation technologies and localization systems into translation studies. They need to consider the changes within the translation industry. The study was limited to the study of translation quality in the Arabic versions of the localized LMSs. The localization quality of other software programs, games, websites, and applications needs to be explored. Finally, it is recommended to develop a quality matrix that encompasses all the dimensions and peculiarities of Arabic localization.

Keywords*—Ambiguity; Arabic; language inconsistencies; Learning Management Systems (LMSs); localization quality

I. Introduction

The recent years have witnessed the development of numerous learning technologies including what have come to be known as the Learning Management Systems (LMSs) to address the changing needs of individuals and institutions all over the world. These newly developed learning systems have revolutionized and changed the way we look at knowledge and skill acquisition [1-4]. LMSs are designed to help organizations manage training events, self-paced courses, and blended learning programs. Organizations and academic institutions have realized that these new learning systems are good alternatives for the traditional rigorous and expensive teaching patterns. This may explain the boom in the LMS market over the recent years.

The unprecedented increase in the development of LMSs can be attributed to different reasons including the development of global economies and multinational institutions as well as the changing patterns within teaching and learning. Furthermore, the emergence of the pandemic COVID-19 has posed strong pressures on the developers of e-learning technologies to develop reliable and effective LMSs that can help students and institutions cope up with this dilemma that has changed our lives in drastic ways and represents unique challenges to all institutions around the world [5-8].

According to Verma [5], Covid-19 has brought out the importance of effective and reliable LMSs that can be usefully used to rescue the prevailing dilemmas in continuing education. In the same saying, Holland [9] asserts that e-learning technologies and LMSs are playing a significant role in current times. Many researchers and educators stress that LMSs have been playing an influential role in supporting the education systems and training institutions during the pandemic COVID-19 [6]. According to Raza, et al. [6], learning technologies and LMSs have proved effective in building effective learning environments during this challenging transition in our contemporary history.

Apart from COVID-19, it is argued that the development of LMSs has changed the nature of learning and training [10-13]. According to Stone [14].

Traditionally, Learning Management Systems (LMS) have been designed to deliver, manage, track, and assess learning activities in a formal learning environment. With new forms of communication and content sharing as well as social networking services (both open and closed), a new generation of systems is emerging to facilitate teaching and learning. These new systems are brought into educational institutions to support new teaching and learning environments and emerging social trends as well as to impact the traditional administration and business models.

The increasingly competitive global marketplace for jobs and education has thus posed different challenges for educational institutions to integrate e-learning technologies and LMSs into their systems. In this regard, the education systems in different Arab countries, Alshahrani and Ally [10] argue, have started to use innovative pedagogies best practices in teaching and learning in all education stages to address the needs of learners and to provide maximum flexibility in learning.

In the face of the digital transformation processes and the pandemic COVID-19, schools, universities, and training institutions have started to integrate learning technologies and LMSs into their teaching and learning practices [8, 15-17]. The selection of an appropriate LMS is usually based on different factors including cost, nature of the courses, age and level of the students, and the availability of an Arabic version of the selected LMS. The rationale is that schools, universities, and training institutions tend to ensure that LMSs are delivered in the same way to all users globally, regardless of linguistic differences, many users still have difficulties dealing with the systems in various target languages. In other words, localization, defined as the process of adapting a software product, website, or application to the linguistic, cultural, and technical requirements, needs, and outlook of a target market [18-23], is one main criterion in the selection of appropriate LMSs.

Over the recent years, there is a close relationship between e-learning systems and technologies on one hand and localization on the other hand. Gauld [24] asserts that localization tools have made the process relatively straightforward for both the educators and their language partners. This has resulted in an exponential growth in the eLearning industry that is projected to be worth $331 billion by 2025.

The selection of appropriate LMSs is not always straightforward. Education and training institutions, therefore, have to ensure the localization quality of the Arabic versions of the LMSs. Localization quality is one of the main requirements for the localization process of an LMS. Reliable localization should be based on unambiguous and understandable language, the appropriate language level, standardization of terminology, provision of sufficient context to the translators, and validation of the target text [19, 22, 25].

In order to support individuals and institutions select proper LMSs, reliable and comprehensive evaluation and assessments of LMSs and the way they work should be generated. In light of this argument, this study seeks to provide an evaluation of the Arabic versions of LMSs. The purpose is to help individuals and institutions with the selection of appropriate LMSs on one hand, and to provide working and reliable solutions and strategies to localization developers to improve their effectiveness and performance.

The rest of this article is organized as follows. Section 2 is Literature Review. It provides a brief survey of localization evaluation and assessment systems. Section 3 is Methodology. It describes the methods and procedures of the study. Section 4 is Analysis and Discussions. It provides both qualitative and qualitative analyses of the data and information gathered through the survey. Section 5 is the Conclusion. It summarizes the main findings, recommendations, and implications of the study.

II. Literature Review

Numerous studies have been recently developed to evaluate and assess the accuracy and reliability of localization systems in different languages, including Chinese, Hindu, Russian, and Spanish. This can be attributed to the growing popularity of localization systems and services which have been parallel to the unprecedented development in web applications and software programs. Zhang [26] comments that currently localization is an important requirement in many international corporations. It is utilized in many kinds of services, such as translation, QA, DTP, testing, and project management. He adds that within the global changes we witness today, localization organizations are providing localization solutions for multiple languages. In this regard, Zhang suggests that it is important to develop evaluation systems to manage localization projects and ensure effective quality management. Likewise, Hassan [27] explains that evaluation systems help both users and localization project managers be confident of the quality delivered. It is universally accepted that localization quality is one of the key success factors of brands, applications, and software programs [28-31].

Given the importance of developing evaluation systems and criteria for localization quality assurance, different approaches have been generated. These approaches have been primarily concerned with assessing the localization quality in terms of consistency and precision of localization systems by contrasting the source and target segments [19, 32-35]. Localization quality, Lobanov [36] defines, is a state of the translated text, such that the language is correct, accurately reflects the idea of the original, takes into account cultural specifics, and is easy for the target audience to read and use. In other words, localization quality is achieved when the translated text reveals the idea of the original accurately in the target language.

The evaluation of localization quality is usually based on three criteria: language translation or linguistic properties, the transition of the product, and the outcome of the product. Allen, et al. [34] asserts that the quality of the translation is one of the main criteria of the localization quality. The quality of the translation entails also defining the target audience. That is, the quality of translation requires that the language is accessible to the target audience. Localization developers should consider the cultural and religious sensitivity of the target audience [18, 20]. Failing to consider such sensitive issues can have negative impacts on the accessibility, marketing, and value of software products and applications [37-39].

Concerning the localization of LMSs, Núñez [40] argues that LMSs should have multilingual support. That is, they should come in different languages because not all students, trainees, and instructors will be able to understand how to navigate the user interface, if it is not translated in their native languages. She asserts that LMSs that support multiple languages for their users, become highly engaging systems as they help optimize user experience by localizing the process

and putting an end to geographical boundaries. In this sense, localization quality is one of the key elements of successful and reliable LMSs.

Gauld [24] argues that the localization processes of LMSs should consider the translation of the written content, graphics, navigation buttons, images, audio and video materials, and data formats. Fadil and Khaldi [41] add that the translation of the contents should be converted into accessible language so that LMSs can make a real impact on learning environments.

Despite the existence of prolific literature on localization evaluation and assessment, research on Arabic localization in general and the Arabic localization of LMSs, in particular, is very sparse. This may be attributed in part to the misconceptions that users in the Arab world generally prefer to use the applications and software programs in their original languages [42, 43]. These misconceptions can be due to the lack of statistics and field studies on linguistic preferences and attitudes towards the use of software applications.

Very few studies, however, have been recently done on evaluating the Arabic localization of some applications, software programs, and games. For instance, Omar and Alqahtani [44] addressed the linguistic challenges in the localization of Enterprise Resource Planning (ERP) systems. The authors reported different linguistic problems related to the use of these systems in the Saudi universities. The authors suggested that all ERP instructions, data, applications, and screens should be made available in a clear and accessible language for the successful implementation of these systems. They finally recommended institutions select the ERP systems that are embedded with multi-language capabilities to address the linguistic needs of all employees and stakeholders. Despite the focus of the study on the language problems and challenges in the Saudi universities and the proper selection of localization systems, it did not provide solutions to Arabic localization problems. Furthermore, the issue of LMSs was not considered within the ERP theoretical framework.

To our knowledge, the first attempt to evaluate the performance of Arabic localizers and the attitudes and perceptions of individuals towards Arabic localization was developed by Al-Mazrooa [42]. In her study of the translation performance in the Arabic versions of some applications including Blackboard Learn and FIFA Video Games, Al-Mazrooa referred to many problems including language inconsistencies that pose serious challenges for users in the Arab countries. Accordingly, she recommends localization developers consider these problems and challenges so that Arab users can use these applications in the same way native users do in their languages. Despite the contributions of the study to the Arabic localization literature, the study was not focused on LMSs. It was confined to Blackboard Learn in the Saudi universities.

In a recent study, however, Omar, et al. [43] attempted a comprehensive analysis of language problems in LMS. In their analysis of the ambiguity in Arabic localization, they focused on LMSs including Blackboard Learn, Microsoft Teams, and Zoom. This study has addressed the issue of localization quality in LMSs through investigating whether the language content and translated texts are clearly conveyed in Arabic. The authors indicate that localization quality is a key factor for the reliability, and successful and effective use of LMSs which requires that language content be consistently converted into Arabic. They also indicated that most users generally prefer to use software applications in their native languages. They reported that in Arab countries, users still prefer to use software applications and LMSs in Arabic despite the dominance of English as a global language. They concluded that ambiguity and linguistic inconsistencies are two serious problems that have negative impacts on the localization quality of LMSs, and thus have adverse impacts on the effectiveness and reliability of such systems. They finally suggested that localization developers pay attention to the peculiar linguistic features of Arabic and develop effective strategies for terminology management.

One main problem with their study, however, is that the results cannot be appropriately generalized. The number of the participants in the study is very limited compared to the recent reports indicating the number of LMSs users in the Arab world. In light of this limitation, this study seeks to bridge the gap in the literature through evaluating the localization quality of the Arabic versions of the LMSs.

III. Methods, Procedures and Results

To explore users' perceptions towards the Arabic versions of the LMSs, an online questionnaire was conducted. The rationale is that questionnaires are appropriate tools for gathering and collecting large amounts of information from a large number of people. The underlying principle was that an online questionnaire is convenient and can be usefully used to bring in a high response, giving also the respondents the flexibility to answer the questions on their own schedule at a pace they choose.

As the study was essentially concerned with the Arabic localization of LMSs, it was appropriate to target only those involved in academic and learning contexts. Participants were asked about their familiarity with LMSs and whether they used the Arabic versions of these systems. They were also asked about their experiences with the Arabic localization of these systems and whether they faced any problems in dealing with the Arabic version.

In total, 7263 participants from 13 Arab countries, including Algeria, Bahrain, Egypt, Jordan, Kuwait, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Tunisia, and the United Arab Emirates, responded to the survey. The participants reflected different age groups, social backgrounds, and nationalities.

In terms of countries, Saudi Arabia and Egypt represent the two biggest participating countries, as shown in Fig. 1. This can be related to the number of internet users, schools, and universities in the two countries. According to reports released by the Egyptian Central Agency for Public Mobilization and Statistics in November 2020, there are about 3.1 million students in the universities and other Higher Education institutions. Concerning Saudi Arabia, the recent years have witnessed the development of many universities. Today, there are 51 universities in Saudi Arabia. Almost all these countries

have adopted digital transformation initiatives over the last five years to convert traditional classes into smart ones [45].

In terms of age, the majority of the participants belong to the age group 18-45, as shown in Fig. 2. This is not surprising anyway. According to Johnson [46], around two-thirds of online users worldwide are aged between 18 and 45 years. It is assumed that these rates do not represent any problems with the representativeness of the data.

The participants included pre-university and university students, teachers, faculty, and information technology (IT) staff in both public and private institutions, as shown in Fig. 3. The participants represented more than 20 universities including King Abdulaziz University, King Saud University, Prince Sattam Bin Abdulaziz University, Northern Borders University, Mansoura University, Assuit University, Port Said University, Suez Canal University, Bahrain University, the University of Jordan, and the University of Nizwa.

Responding to the use and familiarity of LMSs, around 87 of the participants indicated that they used LMSs in their education and training. This is not surprising, however. Almost all educational institutions in the Arab countries have integrated learning technologies and LMSs into their teaching processes over the last two years. This also explains the reason that the majority (around 78% of the participants) used LMSs only over the last two years. This also reflects the fact that LMSs were not widely used before the outbreak of the pandemic COVID-19.

Concerning the most common LMSs, the majority of the participants referred to Blackboard Learn, Blink, Moodle, Microsoft Teams, and Zoom, as shown in Fig. 4.

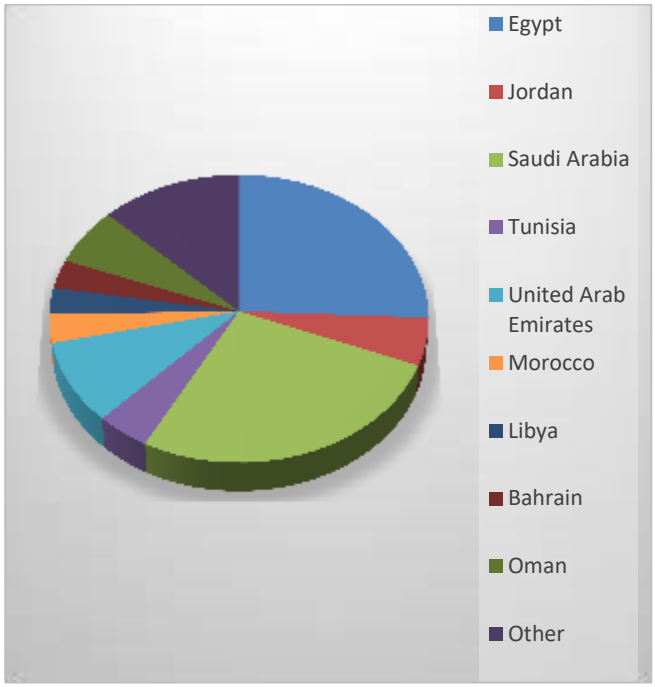


Fig. 1. Participants by Country.

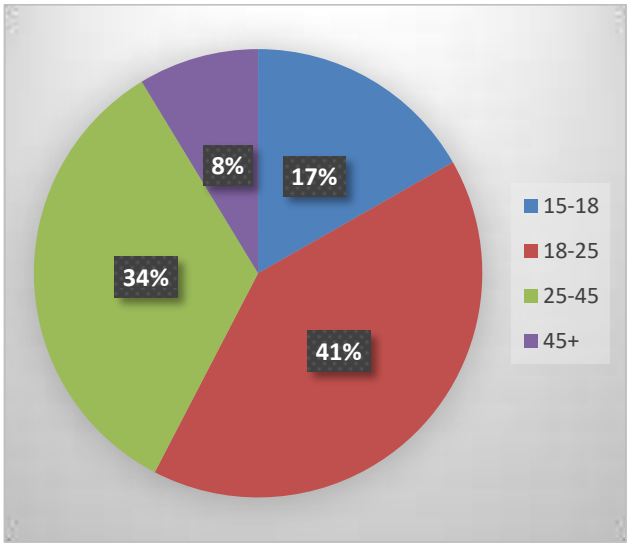


Fig. 2. Participants by Age.

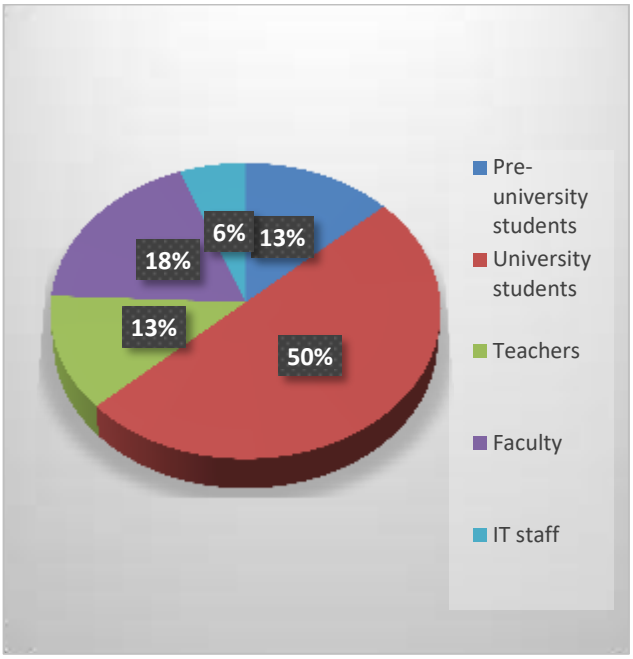


Fig. 3. Participants by Profession.

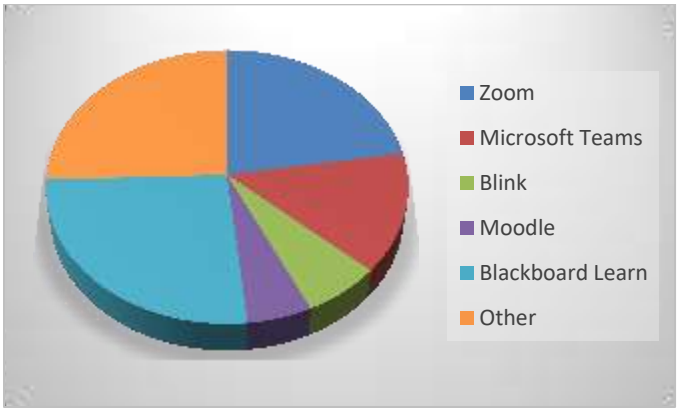


Fig. 4. The Most Commonly used LMSs.

Some of the participants referred to Facebook and WhatsApp as LMSs. Regardless of the recognition of these applications as LMSs, it was clear that many institutions were not ready to cope up with the changes brought up by COVID-19. Instructors and faculty members, therefore, had to use such applications as free and available avenues and channels to communicate with their students [47-49].

73 percent of the participants chose to use the LMS in Arabic with respect to language preferences. However, about 19% indicated that when they had trouble understanding the material in Arabic, they typically moved to the English or original versions. Finally, around 83% stressed that they were facing problems with the use of the Arabic versions of localized LMSs. These problems included ambiguous content, unclear instructions, and language inconsistencies.

IV. Analysis and Discussions

The findings clearly show the widespread of LMSs in the Arab academic institutions as reflected in the participants' familiarity with these systems. The widespread of such systems will definitely have positive impacts on the digital transformation processes and learning processes in the Arab educational institutions. Lack of localization quality, however, is a serious problem that may have adverse impacts on the reliability and effectiveness of such systems.

The major problem with the Arabic versions of the localized LMSs can be described under the heading 'translation inconsistencies'. These inconsistencies definitely have negative impacts on localization quality and reliability [50]. Generally, it is not easy for LMSs users to understand the instructions and terminologies in these systems.

The issue of translation inconsistencies within LMSs can be discussed in the wider context of the challenges of technical translation in Arabic. Technical texts usually contain tough words and expressions which require even tougher translation standards. The translation of technical terms into an accessible language is usually challenging in Arabic [51]. According to Omar, et al. [43], there are thousands of English and Western technical terms that have no direct equivalents in Arabic. It is not easy therefore for translators to convert the terms and instructions in a clear language taking into consideration other factors including space, format, and style. In other words, LMSs usually support different languages where translators are required to be committed to definite templates and designs. Given the fact that not all languages are identical, localization developers, therefore, need to address the linguistic challenges and peculiarities of Arabic, including its morphology, writing system, and dialectal diversity. Indeed, the peculiar linguistic features of Arabic remain among the most serious challenges that have negative impacts on the NPL applications including information retrieval, machine translation, and definitely localization [52-56].

The Arabic localization systems should be aligned with the universal quality standards. These include the quality matrix developed by Lommel, et al. [57]. The matrix is defined into major dimensions including accuracy, design, fluency, internationalization, locale convention, style, terminology, verity, and compatibility. The matrix is graphically represented as shown in Fig. 5.

Finally, the findings of the study indicate clearly that there are serious problems with the translation programs in the Arab universities. Program designers should consider the importance of making students familiar with language technologies and localization systems [58]. Universities and training institutions are also required to consider the changes within the translation industry. Translation should be considered as a product and accordingly, graduates should be prepared and qualified for their careers to stand out in the highly competitive labor markets [59].

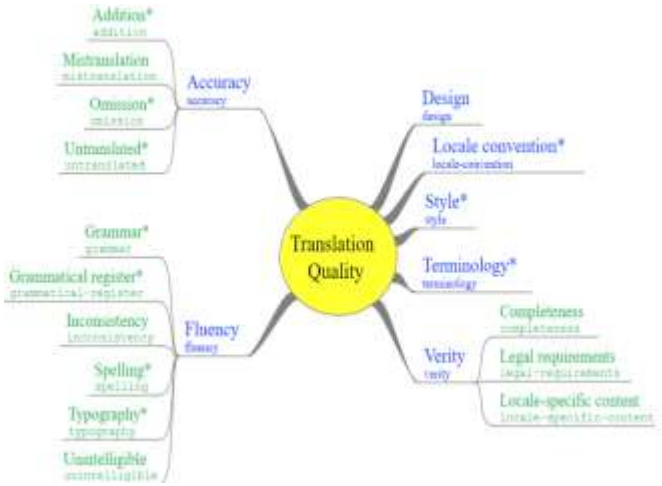


Fig. 5. Quality Matrix Developed by Lommel, et al. [57].

V. Conclusion

There is general dissatisfaction among users of the Arabic versions of the LMSs about the localization quality. The majority of the participants asserted that the use of the English versions of the LMSs is more convenient and reliable. The Arabic versions of LMSs including Blackboard Learn, Zoom, and Microsoft Teams were blamed for ambiguous instructions and language/translation inconsistencies. These problems have negative impacts on the effectiveness and reliability of these systems in schools, universities, and training institutions.

In order to implement LMSs successfully in schools, universities, and training institutions, both localization and translation should go hand-in-hand. Language content, including words, terms, and phrases, should be consistently converted into accessible Arabic. In this sense, localization developers and LMSs designers need to consider the linguistic variations and diversity of Arabic as well as the different needs of users all over the Arab countries. They need also to consider the peculiar linguistic features of Arabic.

The findings of the study have implications for translation programs in Arab universities and training institutions. Program designers should integrate translation technologies and localization systems into translation studies. They need to consider the changes within the translation industry.

The study was limited to the study of translation quality in the Arabic versions of the localized LMSs. The localization quality of other software programs, games, websites, and

applications needs to be explored. Finally, it is recommended to develop a quality matrix that encompasses all the dimensions and peculiarities of Arabic localization.

ACKNOWLEDGMENT

We take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

REFERENCES

[1] Y. Kats, Learning Management System Technologies and Software Solutions for Online Teaching: Tools and Applications: Tools and Applications. Pennsylvania: Information Science Reference, 2010.

[2] Y. Kats, Learning Management Systems and Instructional Design: Best Practices in Online Education. Hershey, Pennsylvania: IGI Global, 2013.

[3] S. B. Dias, J. A. Diniz, and L. J. Hadjileontiadis, Towards an Intelligent Learning Management System Under Blended Learning: Trends, Profiles and Modeling Perspectives. New York: Springer International Publishing, 2013.

[4] S. Foreman, The LMS Guidebook: Learning Management Systems Demystified American Society for Training & Development, 2017.

[5] M. C. Verma, New Paradigm in eLearning Technologies Arising Due To Covid-19 Crisis. EPFRA, 2020.

[6] S. A. Raza, W. Qazi, K. A. Khan, and J. Salam, "Social Isolation and Acceptance of the Learning Management System (LMS) in the time of COVID-19 Pandemic: An Expansion of the UTAUT Model," Journal of Educational Computing Research, p. 0735633120960421, 2020.

[7] Z. Alrefaie, M. Hassanien, and A. Al-Hayani, "Monitoring online learning during COVID-19 pandemic; Suggested online learning portfolio (COVID-19 OLP)," MedEdPublish, vol. 9, 2020.

[8] G. Korkmaz and Ç. Toraman, "Are we ready for the post-COVID-19 educational practice? An investigation into what educators think as to online learning," International Journal of Technology in Education and Science (IJTES), vol. 4, no. 4, pp. 293-309, 2020.

[9] B. Holland, Handbook of Research on Library Response to the COVID-19 Pandemic. Hershey, Pennsylvania: IGI Global, 2021.

[10] K. Alshahrani and M. Ally, Transforming Education in the Gulf Region: Emerging Learning Technologies and Innovative Pedagogy for the 21st Century. London; New York: Routledge, 2017.

[11] S. S. Binyamin, M. J. Rutter, and S. Smith, "Extending the Technology Acceptance Model to Understand Students' use of Learning Management Systems in Saudi Higher Education," International Journal of Emerging Technologies in Learning, vol. 14, no. 3, 2019.

[12] A. I. Saroia and S. Gao, "Investigating university students' intention to use mobile learning management systems in Sweden," Innovations in Education and Teaching International, vol. 56, no. 5, pp. 569-580, 2019.

[13] Y. Tjong, L. Sugandi, A. Nurshafita, Y. Magdalena, C. Evelyn, and N. S. Yosieto, "User Satisfaction Factors on Learning Management Systems Usage," in 2018 International Conference on Information Management and Technology (ICIMTech), 2018, pp. 11-14: IEEE.

[14] D. E. Stone, "Learning Management Systems in a Changing Environment," in Handbook of Research on Education and Technology in a Changing Society, V. X. Wang, Ed. Hershey, Pennsylvania: IGI Global, 2014, pp. 756-767.

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

[16] J. Delcker and D. Ifenthaler, "Teachers' perspective on school development at German vocational schools during the Covid-19 pandemic," Technology, Pedagogy and Education, pp. 1-15, 2020.

[17] C. Carrillo and M. A. Flores, "COVID-19 and teacher education: a literature review of online teaching and learning practices," European Journal of Teacher Education, vol. 43, no. 4, pp. 466-487, 2020.

[18] R. Čermák and Z. Smutny, "A Framework for Cultural Localization of Websites and for Improving Their Commercial Utilization," in Global Observations of the Influence of Culture on Consumer Buying BehaviorEdition: Advances in Business Strategy and Competitive Advantage, S. Sarma, Ed. Hershey, Pennsylvania: IGI Global, 2018.

[19] M. A. Jimenez-Crespo, Translation and Web Localization. London; New York: Routledge, 2013.

[20] K. Keniston, Software Localization: Notes on Technology and Culture. Cambridge: Massachusetts: The Massachusetts Institute of Technology, 1997.

[21] G. Mao and B. Fidan, Localization Algorithms and Strategies for Wireless Sensor Networks: Monitoring and Surveillance Techniques for Target Tracking: Monitoring and Surveillance Techniques for Target Tracking Information Science Reference, 2009.

[22] K. J. Dunne, Perspectives on Localization. J. Benjamins Publishing Company, 2006.

[23] B. Maylath and K. S. Amant, Translation and Localization: A Guide for Technical and Professional Communicators. London; New York: Routledge, 2019.

[24] N. Gauld. (2018, April 15, 2018) Five Things You Need To Know About eLearning Localization. E-Learning Industry. Available: https://elearningindustry.com/elearning-localization-5-things-need-know.

[25] N. Skoryukina, J. Shemiakina, V. L. Arlazarov, and I. Faradjev, "Document localization algorithms based on feature points and straight lines," in Tenth International Conference on Machine Vision (ICMV 2017), 2018, vol. 10696, p. 106961H: International Society for Optics and Photonics.

[26] K. Zhang, "Translation Quality Management Model for Multilanguage localization in Outsourcing Environment," Translation Directory.

[27] M. Hassan. (2018, March 22, 2018) Translation Quality Assessment: Error Categories and Severity. Vocalink Global. Available: https://vocalinkglobal.com/translation-quality-assessment-errors-categories/.

[28] M. Nogueira Cortimiglia, A. Ghezzi, and F. Renga, "Social applications: Revenue models, delivery channels, and critical success factors-An exploratory study and evidence from the Spanish-speaking market," Journal of theoretical and applied electronic commerce research, vol. 6, no. 2, pp. 108-122, 2011.

[29] V. R. Tummala, C. L. Phillips, and M. Johnson, "Assessing supply chain management success factors: a case study," Supply Chain Management: An International Journal, 2006.

[30] R. Heeks and B. Nicholson, "Software export success factors and strategies in 'follower'nations," Competition & Change, vol. 8, no. 3, pp. 267-303, 2004.

[31] S. Shin, H. Kim, and W. Kim, "Transnational corporations' localization strategies via retail attributes: Focus on Chinese Market," Journal of Retailing and Consumer Services, vol. 55, p. 102088, 2020.

[32] B. Esselink, A Practical Guide to Localization. John Benjamins Publishing Company, 2000.

[33] M. C. Chao, N. Singh, and Y. N. Chen, "Web site localization in the Chinese market," Journal of Electronic Commerce Research, vol. 13, no. 1, p. 33, 2012.

[34] M. Allen, S. Baydere, E. Gaura, and G. Kucuk, "Evaluation of localization algorithms," in Localization Algorithms and Strategies for Wireless Sensor Networks: Monitoring and Surveillance Techniques for Target TrackingHershey, Pennsylvania: IGI Global, 2009, pp. 348-379.

[35] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2568-2599, 2019.

[36] M. Lobanov. (2014, August 21, 2014) Localization quality. A myth or reality? Multilingual. 1-4.

[37] S. Abufardeh and K. Magel, "The impact of global software cultural and linguistic aspects on Global Software Development process (GSD): Issues and Challenges," in 4th International Conference on New Trends in Information Science and Service Science, 2010, pp. 133-138: IEEE.

[38] G. Ger, "Localizing in the global village: Local firms competing in global markets," California Management Review, vol. 41, no. 4, pp. 64-83, 1999.

[39] S. A. Becker, "An exploratory study on web usability and the internationalization of US e-businesses," J. Electron. Commerce Res., vol. 3, no. 4, pp. 265-278, 2002.

[40] M. Núñez. (2019, July 11, 2019) Reasons to Translate Learning Management Systems. SuimulTrans. Available: https://www.simultrans.com/blog/reasons-to-translate-learning-management-systems.

[41] O. A. Fadil and M. Khaldi, "Learning Management Systems: Concept and Challenges," in Personalization and Collaboration in Adaptive E-LearningHershey, Pennsylvania: IGI Global, 2020, pp. 158-175.

[42] N. Al-Mazrooa, Arabic Localisation: Key Case Studies for Translation Studies. Cardiff: Cardiff University, 2018.

[43] A. Omar, I. E. A. W. Shaalan, and W. I. Hamouda, "Ambiguity Resolution in Arabic Localization," (in English), Applied Linguistics Research Journal, vol. 5, no. 1, pp. 1-6, 2021.

[44] A. Omar and M. A. Alqahtani, "The implications of linguistic diversity for the ERP implementation practices in multilingual contexts," International Journal of Advanced and Applied Sciences, vol. 5, no. 7, pp. 46-52, 2018.

[45] A. Omar and A. Almaghthwi, "Towards an Integrated Model of Data Governance and Integration for the Implementation of Digital Transformation Processes in the Saudi Universities," International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, pp. 588-593, 2020.

[46] J. Johnson. (2021). Distribution of internet users worldwide as of 2019, by age group. Available: https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/.

[47] A. E. E. Sobaih, A. M. Hasanein, and A. E. Abu Elnasr, "Responses to COVID-19 in higher education: Social media usage for sustaining formal academic communication in developing countries," Sustainability, vol. 12, no. 16, p. 6520, 2020.

[48] R. A. Machado, P. R. F. Bonan, D. E. d. C. Perez, and H. Martelli JÚnior, "COVID-19 pandemic and the impact on dental education: discussing current and future perspectives," Brazilian oral research, vol. 34, 2020.

[49] N. Ghounane, "Moodle or Social Networks: What Alternative Refuge Is Appropriate to Algerian EFL Students to Learn during COVID-19 Pandemic," Arab World English Journal, vol. 11, no. 3, pp. 21-41, 2020.

[50] M. Lobanov and I. Hill, "Personal brand and localization management " Multilingual, vol. March, pp. 27-31, 2017.

[51] E. M. Muhiesen and M. S. Al-Ajrami, "Challenges in Translating Technical Texts," Dirasat, Human and Social Sciences, vol. 46, no. 1, pp. 322-328, 2019.

[52] A. Omar and M. Aldawsari, "Lexical Ambiguity in Arabic Information Retrieval: The Case of Six Web-Based Search Engines," International Journal of English Linguistics, vol. 10, no. 3, pp. 219-228, 2020.

[53] A. Omar and W. I. Hamouda, "The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic," International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, pp. 116-121, 2020.

[54] A. Omar and W. I. Hamouda, "Document length variation in the vector space clustering of news in arabic: A comparison of methods," International Journal of Advanced Computer Science and Applications (IJACSA), no. 2, pp. 75-80, 2020.

[55] A. Omar, B. I. Elghayesh, and M. A. M. Kassem, "Authorship attribution revisited: The problem of flash fiction a morphological-based linguistic stylometry approach," 2019.

[56] H. N. Alsager, "Towards a Stylometric Authorship Recognition Model for the Social Media Texts in Arabic," 2020.

[57] A. Lommel, A. Burchardt, and H. Uszkoreit, Multidimensional Quality Metrics (MQM) Definition. German Research Center for Artificial Intelligence (DFKI) and QTLaunchPad, 2015.

[58] A. Omar, A. F. Khafaga, and I. E.-N. A. W. Shaalan, "The Impact of Translation Software on Improving the Performance of Translation Majors."

[59] Y. A. Gomaa, R. AbuRaya, and A. Omar, "The Effects of Information Technology and E-Learning Systems on Translation Pedagogy and Productivity of EFL Learners," in 2019 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2019, pp. 1-6: IEEE.

Comparative Analysis of the Impact on Air Quality Due to the Operation of La Oroya Metallurgical Complex using the Grey Clustering Method

Alexi Delgado[1], Luis Vasquez[2], Luis Espinoza[3], Manuel Mejía[4]
Erick Yauri[5], Chiara Carbajal[6], Enrique Lee Huamaní[7]
Mining Engineering Section, Pontificia Universidad Católica del Perú, Lima-Perú[1, 2, 3, 4, 5]
Administration Program, Universidad de Ciencias y Humanidades, Lima-Perú[6]
Image Processing Research Laboratory, Universidad de Ciencias y Humanidades, Lima-Perú[7]

***Abstract*—Air pollution is one of the biggest problems worldwide due to the increase of burning of fossil fuels by industries around the world. In the present work, the air quality study will be carried out with the grey clustering method, since the data obtained presents a certain level of uncertainty. In order to obtain a correct analysis of air quality, the comparison was made in two different years with the same monitoring stations. The air quality assessment was carried out in three monitoring stations located in three different districts of the province of La Oroya (La Oroya Antigua, minor town of Huari and Santa Rosa de Sacco), in which, they installed sampling equipment for the evaluation on the basis of 10 particulate matter (PM10) and sulfur dioxide (SO2). In each point of study a positive result was obtained, where an improvement in air quality can be seen, this it is due to the reduction of mining activity in the study area. These results show the improvement over the years. Finally, this method can also be used by any organization in the nation for water or air quality studies.**

***Keywords*—Air quality assessment; Grey clustering method; particulate matter (PM10); sulfur dioxide (SO2)**

I. Introduction

Air pollution has become a serious problem for large cities due to transportation activities [1] or industrialization [2] as studies show [3], [4]. In this paper, we apply the grey clustering method to assess air quality, which is based on grey systems theory [5], [6]. The grey clustering method can be applied using incidence matrices or weight functions. In this work, air quality will be measured using the center-point triangular whitenization weight functions (CTWF) method because it helps respondents to find good answers at the central point of the intervals called grey classes [7].

In this work, we carry out an evaluation of air quality in the city of La Oroya, Junín, Peru. This city has a massive problem related to air pollution, of great magnitude in some districts, due to various factors, mainly due to mining activity [8], [9]. Therefore, the main objective of this work is the evaluation of air quality at three points in the city of La Oroya, applying the grey scream method. Based on this, a study focused on establishing a comparison and evaluation of air quality according to standard levels between the period of 2007 and 2012 was carried out. Because air quality assessment is a subject with a high level of uncertainty, we firmly believe that applying this method would be of great help since the CTWF method considers uncertainty in its analysis and gives weight to the criteria [10]–[14].

For such purpose, the present investigation is organized as follows. In Section 2, a literature review about the CTWF method is given. Then, in Section 3, the methodology is explained step by step. After that, in Section 4, the results and discussions are presented. Finally, the conclusions are provided in Section 5.

II. Literature Review

Studies by various authors were analyzed in order to obtain useful knowledge for the elaboration and application of the grey clustering method for the evaluation of air quality.

First, it was essential to establish the most important parameters to be used in order to assess air quality. In order to do this we rely on The Association between Air Pollution and Population Health Risk for Respiratory Infection: A Case Study of Shenzhen China [2], in which the probable relationship between exposure to different types of air pollution with respiratory problems in certain periods of time is analyzed. This concludes that there is indeed a relationship between air pollution and respiratory health problems and that the main pollutants were PM10, PM2.5 AND NO2.

Similarly, we have considered the article "Short - term Effects of Ambient Gaseous Pollutants and Particulate Matter on Daily Mortality in Shanghai, China" [15], which analyzes the effect of pollutants and gases on daily mortality in China over time periods. In this study the direct relationship between pollutants PM10, SO2, NO2 with non-accidental mortality and cardiopulmonary diseases in China is verified.

Likewise, statistical information was necessary for the application of the grey clustering method, so the information provided by DIGESA in 2007 and 2012 [16] was used, which consists of the results obtained on the measurement of polluting agents for air, such as PM10, SO2, etc. It is worth mentioning that the data provided by DIGESA have a high degree of reliability, so any result obtained through them will also have a high degree of credibility.

Besides that the variables to be analyzed with the grey clustering method were selected. This selection was based on which variables were most relevant and have the greatest impact in terms of air pollution, in addition to their relation to the elements that originate from the mining industry [17]. The mentioned parameters are used in the grey clustering methodology to obtain a classification of the analyzed parameter.

Finally, the research developed in "Evaluating impact of air pollution on different diseases in Shenzhen, China" [18] was considered as a review for the development of the present paper. Such study assess the adverse health effects of air pollution. This study investigates the excess risk of 6 air pollutants (PM 10, PM 2.5, SO2, NO2 ...) for 21 disease groups. Daily air quality data and 1.6 million outpatient visit records from Shenzhen, China are used in the study. Where the results show that associations between air pollutants and diseases vary across different disease groups.

III. Methodology

The methodology presented in this work, to compare the degree of contamination, is the grey clustering.

To use this, we first define a set of "n" criteria (our classification of data to compare), a set of "s" grey classes (good, fair, bad, very bad) and a set of "m" objects (study points), according to sample values 〖 x〗 _ij. The steps of the method are as follows:

Step 1: The center-points λ1,λ2,…,λs of the grey classes are defined.

Step 2: Triangular functions are based on the relation to the number of "n" grey classes, which allows us to obtain the correspondence rule for triangular functions as represented in Fig. 1.

For a $f_j^k(x_{ij})$ calculated by (1).

$$f_j^k(x_{ij}) = \begin{cases} \frac{x-\lambda_2}{\lambda_3-\lambda_2}, & x \in [\lambda_2, \lambda_3] \\ \frac{\lambda_n-x}{\lambda_n-\lambda_3}, & x \in <\lambda_3, \lambda_n> \\ 0, & x \notin [\lambda_2, \lambda_n] \end{cases} \quad (1)$$

Step 3: The comprehensive clustering coefficient σ_i^k for object $i, i = 1, 2, \dots, m$, in the grey class $k, k = 1, 2, \dots, s$, is defined by (2).

$$\sigma_i^k = \sum_{j=1}^{n} f_j^k(x_{ij}) \cdot n_j \quad (2)$$

Where $f_j^k(x_{ij})$ is the CTWF, and η_j is the weight of criterion j.

Step 4: If $\max_{1 \le k \le s} \{\sigma_i^k\} = \sigma_i^{k^*}$, it is decided that object i belongs to grey class $k*$. If there are several objects in grey class $k*$, these objects could be ordered according to the values of σ_i^k.

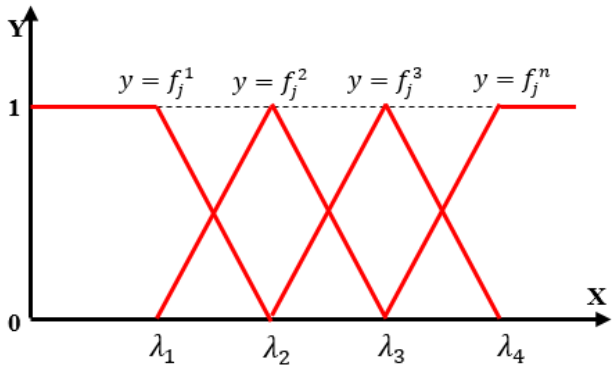


Fig. 1. Triangular Function.

IV. Case Study

A. Monitoring Points

This study is based on the evaluation of air quality, which includes the La Oroya district as a scope, where three fixed stations were established in the following areas: Oroya Antigua, minor town of Huari and Santa Rosa de Sacco, installing equipment sampling for particle evaluation [19] as illustrated in Fig. 2.

Fig. 2. Location of Monitoring Points.

The monitoring points along with their locations and codes can be found in Table I.

TABLE I. Monitoring Points

code	Monitoring point	Location	District	Altitude	Coordinates	
					East	North
P1	INSTITUCIÓN EDUCATIVA NACIONAL N° 31149	Av. Brasil N° 222 minor town of Huari	La Oroya	3682 m	409394	8712744
P2	Vivienda	St. Dos de Mayo	Oroya Antigua	3728 m	401953	8726184
P3	Municipalidad Distrital de Santa Rosa de Sacco	Street Mariano Melgar N° 208	Santa Rosa de Sacco	3789 m	397482	8723112

B. Evaluation Parameters

For the evaluation of air quality, the Peruvian law [20] establishes four levels of air quality for each parameter. The parameter values are shown in Table II.

Likewise the data of each parameter collected in 2007 is shown in Table III.

Besides, the data of each parameter collected in 2012 is shown in Table IV.

Then, the dimensioned data for each monitoring point during 2007 is displayed in Table V.

The same data regarding the year of 2012 is given in Table VI.

Then the steps of the methodology according to Section 3 are applied.

TABLE II. AIR QUALITY LEVELS

Levels		**Parameters**	
		C1 (PM10)	***C2 (SO2)***
λ1	Good	"0 -75"	"0 -10"
λ2	Moderate	"76 - 150"	"11 - 20"
λ3	Bad	"151 - 250"	"21 - 500"
λ4	Care threshold	> 250	> 500

TABLE III. DATA OF EACH PARAMETER FOR EACH MONITORING POINT IN 2007

2007	**P1**	**P2**	**P3**
C1	42.44	44.92	70.74
C2	287.92	1237.27	64.41

TABLE IV. DATA OF EACH PARAMETER FOR EACH MONITORING POINT IN 2012

2012	**P1**	**P2**	**P3**
C1	38.7	28.4	48.5
C2	11	33	7.66

TABLE V. DIMENSIONED DATA 2007

2007	**P1**	**P2**	**P3**
C1	0.266	0.282	0.444
C2	1.467	6.305	0.328

TABLE VI. DIMENSIONED DATA 2012

2012	**P1**	**P2**	**P3**
C1	0.243	0.178	0.304
C2	0.056	0.168	0.039

Step 1: From Table II, of the air levels, the mean and dimensioned points were determined for each value of λ, this is shown in Table VII.

TABLE VII. DIMENSIONED PARAMETERS

	λ1	**λ2**	**λ3**	**λ4**
C1	0.235	0.706	1.255	1.804
C2	0.025	0.076	1.325	2.573

Step 2: From Table VII, the triangular functions [21] for each parameter C1 and C2 will be obtained.

As an example, these functions will be calculated for a single monitoring point as shown in (3) – (10); in which (3) – (6) are a demonstration regarding the year of 2007, while (7) – (10) displays the data of the year 2012.

$$f_j^1(x) = \begin{cases} 1, & x \in [0,0.235] \\ \frac{0.706-x}{0.706-0.235}, & x \in <0.235,0.706> \\ 0, & x \in [0.706,\infty> \end{cases} \quad (3)$$

$$f_j^2(x) = \begin{cases} \frac{x-0.235}{0.706-0.235}, & x \in [0.235,0.706], \\ \frac{1.255-x}{1.255-0.706}, & x \in <0.706,1.255> \\ 0, & x \notin [0.235,1.255] \end{cases} \quad (4)$$

$$f_j^3(x) = \begin{cases} \frac{x-0.706}{1.255-0.706}, & x \in [0.706,1.255], \\ \frac{1.804-x}{1.804-1.255}, & x \in <1.255,1.804> \\ 0, & x \notin [0.706,1.804] \end{cases} \quad (5)$$

$$f_j^4(x) = \begin{cases} \frac{x-1.255}{1.804-1.255}, & x \in [1.255,1.804], \\ 1, & x \in <1.804,\infty> \\ 0, & x \in [0,1.255> \end{cases} \quad (6)$$

$$f_j^1(x) = \begin{cases} 1, & x \in [0,0.025] \\ \frac{0.076-x}{0.076-0.025}, & x \in <0.025,0.076> \\ 0, & x \in [0.076,\infty> \end{cases} \quad (7)$$

$$f_j^2(x) = \begin{cases} \frac{x-0.025}{0.076-0.025}, & x \in [0.025,0.076], \\ \frac{1.325-x}{1.325-0.076}, & x \in <0.076,1.325> \\ 0, & x \notin [0.025,1.325] \end{cases} \quad (8)$$

$$f_j^3(x) = \begin{cases} \frac{x-0.076}{1.325-0.076}, & x \in [0.076,1.325], \\ \frac{2.573-x}{2.573-1.325}, & x \in <1.325,2.573> \\ 0, & x \notin [0.076,2.573] \end{cases} \quad (9)$$

$$f_j^4(x) = \begin{cases} \frac{x-1.325}{2.573-1.325}, & x \in [1.325,2.573], \\ 1, & x \in <2.573,\infty> \\ 0, & x \in [0,1.325> \end{cases} \quad (10)$$

The values in Table VII were replaced in the triangular functions to calculate the CTWF values in each criterion for both years of evaluation. Such results can be found in Table VIII and Table IX.

Step 3: To find the clustering coefficient, the CTWF values in Tables VIII and Table IX must be multiplied by a criterion weight. This weight is a function of the inverses of the dimensioned values in Table VII, such weights are given in Table X.

The values of the coefficient obtained are presented in Table XI and Table XII, in which the maximum value is highlighted in yellow in order to emphasize it.

Step 4: Finally, the classification of each sampling point is made according to the maximum clustering coefficient values for each year, obtaining the following:

TABLE VIII. CTWF VALUES FOR EACH MONITORING POINT FOR 2007

P1	**C1**	**C2**
f1	0.9348	0.0000
f2	0.0653	0.0000
f3	0.0000	0.8861
f4	0.0000	0.1139
P2	**C1**	**C2**
f1	0.9008	0.0000
f2	0.0992	0.0000
f3	0.0000	0.0000
f4	0.0000	1.0000
P3	**C1**	**C2**
f1	0.5565	0.0000
f2	0.4435	0.7985
f3	0.0000	0.2015
f4	0.0000	0.0000

TABLE IX. CTWF VALUES FOR EACH MONITORING POINT FOR 2012

P1	**C1**	**C2**
f1	0.9836	0.4010
f2	0.0164	0.5990
f3	0.0000	0.0000
f4	0.0000	0.0000
P2	**C1**	**C2**
f1	1.0000	0.0000
f2	0.0000	0.9267
f3	0.0000	0.0733
f4	0.0000	0.0000
P3	**C1**	**C2**
f1	0.8540	0.7346
f2	0.1460	0.2654
f3	0.0000	0.0000
f4	0.0000	0.0000

TABLE X. CRITERION WEIGHT FOR EACH PARAMETER

	λ1	**λ2**	**λ3**	**λ4**
C1	0.098	0.098	0.514	0.588
C2	0.902	0.902	0.486	0.412

TABLE XI. VALUES OF THE CLUSTERING COEFFICIENTS FOR 2007

2007	**λ1**	**λ2**	**λ3**	**λ4**
P1	0.091	0.006	0.431	0.047
P2	0.088	0.010	0.000	0.412
P3	0.054	0.764	0.098	0.000

TABLE XII. VALUES OF THE CLUSTERING COEFFICIENTS FOR 2012

2012	λ1	λ2	λ3	λ4
P1	0.458	0.542	0.000	0.000
P2	0.000	0.836	0.036	0.000
P3	0.663	0.239	0.000	0.000

Regarding the year of 2007:

- For P1, $\max_{1\le k\le s}\{\sigma_i^k\} = 0.431$, where k=3. Therefore, P1 belongs to bad grey class.
- For P2, $\max_{1\le k\le s}\{\sigma_i^k\} = 0.412$, where k=4. Therefore, P2 belongs to care threshold grey class.
- For P3, $\max_{1\le k\le s}\{\sigma_i^k\} = 0.764$, where k=2. Therefore, P3 belongs to moderate grey class.

Regarding the year of 2012:

- For P1, $\max_{1\le k\le s}\{\sigma_i^k\} = 0.542$, where k=2. Therefore, P1 belongs to moderate grey class
- For P2, $\max_{1\le k\le s}\{\sigma_i^k\} = 0.836$, where k=2. Therefore, P2 belongs to moderate grey class
- For P3, $\max_{1\le k\le s}\{\sigma_i^k\} = 0.663$, where k=1. Therefore, P3 belongs to good grey class

V. RESULT AND DISCUSSION

A. About the Case Study

In this section, the results obtained regarding the air quality of the city of La Oroya and its relationship with the activity of the city's metallurgical complex will be presented.

Regarding the results obtained in 2007, it was observed from the three monitoring points that none of them registered good air quality, in accordance with the air quality incidences of the Peruvian state [22].

Likewise, it is possible to establish that the three control points have different air qualities despite being in the same province. Thus, it is established that control point P3 presents better air quality than points P2 and P1, and that control point P1 presents better air quality than point P2, obtaining the following relation:

P3>P1>P2

On the other hand, regarding to the results achieved in 2012, it was observed that the P3 monitoring point recorded good air quality and, as for the other two points, air quality was improved, compared to 2007 [16].

It is also possible to establish that the three control points have different air qualities, which provides that the P3 has a better air quality than points P2 and P1. Following the order registered in 2007.

In a comparison between the results obtained in 2007 and 2012, we can see an improvement in air quality, which can be explained by the closure of the metallurgical complex La Oroya between June 2009 and August 2012, the month in which the zinc plant resumes its operations [23].

Finally, if a comparison is made between the distances of the control points and the metallurgical complex, it is observed that the tracking point P1 is 15 km away, P2 is 0.5 km and P3 5 km away; however, point P1 records a worse air quality than point P3, this is due to the direction of the air currents that are recorded in the area.

B. About the Methodology

In relation to the method, grey clustering, was applied in the analysis of 10 particulate matter (PM10) and sulfur dioxide (SO2), with which it was possible to obtain clusterization equations and assign weights to each of the pollutants, which allows us to establish the quality of air from the city of La Oroya with respect to each pollutant analyzed.

It should be noted that in this work the weights were assigned to the specified pollutants starting from the use of the grey clustering method; however, it is necessary to establish an in-depth evaluation of the weight corresponding more precisely to each pollutant.

VI. Conclusions

The air quality in 2012 compared to 2007 shows an improvement in all the evaluated control points, which may be due to the fact that in this period the mining company entered into a total temporary freeze due to social and environmental conflicts. Likewise, it was also observed that the control points closer to the metallurgical complex did not necessarily record lower air quality relative to the more distant monitoring points, this may be due to the trajectory of air currents in the area.

It also has been demonstrated that along with the grey clustering and CTWF method an air quality analysis can be made based on specific points or on different points in the time series throughout the year, since it is based on the theory of the grey system, which analyzes the uncertainty and therefore its use in the evaluation of air quality is appropriate due to the high uncertainty that it possesses.

The method used in this work can be considered as an appropriate option for the evaluation of nitrogen dioxide in the air, since no continuous record of this contaminant has been found by DIGESA, in the period that this research was carried out. On the other hand, it is suggested to analyze the possible relationship between respiratory infections and the various air pollutants in La Oroya. Finally, it is recommended to analyze the air quality in the cities surrounding the city of La Oroya and its possible linkage with the metallurgical complex in future studies for a better understanding of the topic and its evolution in this context.

References

[1] V. S. Limaye et al., Development of ahmedabad's air information and response (Air) plan to protect public health, Int. J. Environ. Res. Public Health, vol. 15, no. 7, Jul. 2018.

[2] X. Xia, A. Zhang, S. Liang, Q. Qi, L. Jiang, and Y. Ye, The association between air pollution and population health risk for respiratory infection: A case study of Shenzhen, China, Int. J. Environ. Res. Public Health, vol. 14, no. 9, Sep. 2017.

[3] A. Delgado and A. Aguirre, Air Quality level Assessment through the Grey Clustering Analysis on Lima, Peru. .

[4] A. Delgado, P. Montellanos, and J. Llave, Air quality level assessment in Lima city using the grey clustering method, in IEEE ICA-ACCA 2018 - IEEE International Conference on Automation/23rd Congress of the Chilean Association of Automatic Control: Towards an Industry 4.0 - Proceedings, 2019.

[5] S. Liu and Y. Lin, Grey Systems Theory and Applications, vol. 53. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[6] S. Liu and Y. Lin, Introduction to grey systems theory, vol. 68. 2010.

[7] A. Delgado, A. Espinoza, P. Quispe, P. Valverde, and C. Carbajal, Water quality in areas surrounding mining: Las Bambas, Peru, Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, 2019.

[8] El Comercio Perú, La Oroya: todo lo que tienes que saber sobre el conflicto, El Comercio Perú, Junín, 12-Aug-2015.

[9] A. Delgado, R. Cuba, H. Jamanca, A. Sampen, and C. Carbajal, Social impact assessment in la oroya, peru applying the grey clustering method, Int. J. Innov. Technol. Explor. Eng., vol. 9, no. 1, 2019.

[10] A. Delgado, H. Reyes, I. Romero, and C. C. Mancilla, Social impact assessment using the grey clustering method: A case study on a mining project, 2019, pp. 1–5.

[11] A. Delgado, E. Luna, M. Hernández, K. Montero, and C. Carbajal, Assessment of the air quality in four cities with near mining activity in mexico, using the grey clustering method, Int. J. Recent Technol. Eng., vol. 8, no. 3, 2019.

[12] A. Delgado, C. Carbajal, H. Reyes, and I. Romero, Social Impact Assessment on a Mining Project in Peru Using the Grey Clustering Method and the Entropy-Weight Method, Commun. Comput. Inf. Sci., vol. 1096 CCIS, pp. 116–128, 2019.

[13] A. Delgado, N. Rojas, J. Oblitas, B. Andrés, A. Huerta, and C. Carbajal, Water quality assessment using the grey clustering analysis on a river of Taxco, Mexico, Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 4717–4723, 2020.

[14] A. Delgado, J. Culqui, G. Tasayco, A. Millán, E. Tirado, and C. Carbajal, Quality assessment of surface water associated with a copper mine in peruusinggrey systems, Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 6660–6668, 2020.

[15] G. Chen et al., Short-term effects of ambient gaseous pollutants and particulate matter on daily mortality in Shanghai, China, J. Occup. Health, vol. 50, no. 1, pp. 41–47, Jan. 2008.

[16] Ministerio de Salud del Perú, La Oroya, pp. 1–37, 2007.

[17] Y. Zhang, J. Ni, J. Liu, and L. Jian, Grey evaluation empirical study based on center-point triangular whitenization weight function of Jiangsu Province industrial technology innovation strategy alliance, Grey Syst. Theory Appl., vol. 4, no. 1, pp. 124–136, Jan. 2014.

[18] L. L. Chen, J. Xu, Q. Zhang, Q. H. Wang, Y. Q. Xue, and C. R. Ren, Evaluating impact of air pollution on different diseases in Shenzhen, China, IBM J. Res. Dev., vol. 61, no. 6, pp. 21–29, Nov. 2017.

[19] I. Sociales, Anuario de Estadísticas Ambientales, 2012.

[20] D. S. N. 063-2007 PCM, Decreto Supremo, D. Of. El Peru, pp. 1–12, 2008.

[21] L. Zhou and S. Xu, Application of Grey Clustering Method in Eutrophication Evaluation of Wetland, 2007.

[22] Ministerio del Ambiente, Resolución Ministerial N°181-2016-MINAM. Indic de calidad del aire, pp. 1–6, 2016.

[23] Ministerio de Salud, Informe N° 003939-2012/DEPA/DIGESA, Vigil. Sanit. la Calid. del Aire por el reinicio las Act. del Complejo Met. Doe Run Perú, 2012.

Deep Wavelet Neural Network based Robust Text Recognition for Overlapping Characters

Neha Tripathi[1]*, Pushpinder Singh Patheja[2]
School of Computing Science and Engineering
VIT Bhopal, India

***Abstract*—This paper presents a deep learning based intelligent text recognition system with touching and overlapped characters. The robustness and effectiveness in the proposed model are enhanced through the modified configuration of neural network known as Deep Wavelet Neural Network (DWNN). The capability of deep learning networks to learn efficiently from an unlabeled dataset has attracted the attention of many researchers over the last decade. However, the performance of these networks is subject to the quality of the dataset and invariant image representation. Numerous optical character recognition techniques have also been presented in the recent years, but the overlapped and touching characters have not been addressed much. The nonlinear and uncertain representation of image data in case of overlapped text adds severe complexity in the process of feature extraction and respective learning. The proposed architecture of DWNN uses fast decaying wavelet functions as activation function in place of conventional sigmoid function to cope up with the uncertainties and nonlinearity of the data representation in overlapped text images. It comprises of cascaded layered architecture of translated and dilated versions of wavelets as activation functions for the training and feature extraction at multiple levels. The local transformation and deformation variation in the visual data has also been taken care efficiently through the modified architecture of DWNN. Comprehensive experimental analysis has been performed over various test images to verify the effectiveness of the proposed text recognition system. The performance of the proposed method is assessed with the help of the metrics, namely, estimation error, cost function and accuracy. The proposed approach will be implemented in MATLAB.**

Keywords*—Text recognition; overlapped characters; deep wavelet neural network; feature extraction; segmentation; basis function; optical character recognition

I. INTRODUCTION

The field of optical character recognition has attracted a lot of attention over the last two decades due to its capability to extract the meaningful information from the printed or handwritten text. It has been used successfully in the applications like automatic language translation, text to speech converters, smart scanning devices, text summarization, automated postal address and ZIP code reading, bank cheque reading, etc. The characters are recognized through the process of conversion of these characters into the machine-readable formats like ASCII code. The intended information is extracted from the images based on a thorough analysis of the text and graphical features of the document. Both these feature sets are processed to deal with the textual and non-textual component of the image of the document. The non-textual components involve company logos, emoticons, make up line diagrams, delimiting lines between text, etc. The variety and diversity introduced in text extraction through different font sizes, font types, and orientation make the problem of OCR even more challenging [1-3].

A typical framework of OCR involves the process of preprocessing, segmentation, feature extraction and recognition. These processes need to be robust as well as efficient to present an accurate text extraction model. Various techniques have been proposed by numerous researchers to address the problems and challenges faced at these processes. However, segmentation process is the most complicated stage in OCR because of its dependency over large number of factors including the quality of scanner/camera, illumination, font type, font size, orientation, angular features, ink diffusion, etc. [4]. The process of feature extraction and recognition is also an important phase of OCR as it governs the accuracy of the overall process. Various techniques ranging from statistical models to deep learning framework have been proposed for the text recognition based on the characteristics of the features of documents [5-8].

The overall performance of any OCR technique relies majorly upon the quality of image. Presence of noise in the image can greatly degrade the efficiency and accuracy of the process. Although the preprocessing phase is responsible to filter out the noise present in the image, but it is subject to the type of the respective noise. It is also assumed in most of the OCR techniques that the text lines are equitably straight and the distance between neighboring text lines is precise. However, these assumptions could be characterized for text with overlapped characters and slanted orientation. Therefore, the complexity associated with the problem of extracting the text from the images with overlapped or touching characters has attracted the attention over the last few years. But extremely limited work has been done in this field so far. The major problem in designing an efficient OCR framework for overlapped or touched characters is to eliminate the noise from the binary images and smooth them for the feature extraction and recognition. The training algorithm must be intelligent and adaptive enough to deal with the abrupt feature variations generated due to the overlapping [9-12].

Various machine learning and deep learning algorithms have been presented by the researchers for the feature extraction and recognition of text from images. But their accuracy is greatly subject to the textual distribution. Even the best OCR techniques for normal text distribution are found to

*Corresponding Author

perform very poorly for the overlapped characters images. This is because of the reason that the training of the intelligent framework is done through the datasets where the characters are distant and clearly separated from each other [13]. The same network fails to recognize the characters in case of overlapped distribution. Therefore, it requires an adaptive intelligent model which could be able to mitigate the effects of sharp and abrupt changes in the text features distribution.

Some novel configurations of deep neural networks like Convolutional Neural network (CNN), Long Short-term Memory (LSTM)/ Recurrent Neural Network (RNN), Generative Adversarial Network (GAN), etc. have got a lot of appreciation and attention due to their superior learning characteristics and efficient classification performance [14-18]. These networks have also been implemented by various researchers for OCR problems with varying complexity. However, these networks could not be able to present the promising results in highly dynamic and uncertain feature distribution space. The capability of these networks to provide a fast decision making with long term learning and lesser time complexity is limited and therefore the derived model is found to be conservative learning model. The major reason behind this is the basis function used in these networks which is not orthogonal in nature and results into an inefficient and non-unique representation of decision space. Neural networks with non-orthogonal activation functions cannot guarantee the convergence of the learning curve and may get trapped in local minima for certain initial conditions [19].

These limitations of conventional neural network framework have been addressed in some literatures and various novel activation functions have been proposed. However, the replacement of sigmoid functions in DWNN by rapidly decaying functions known as wavelets has generated very promising results for the dynamic and uncertain feature distribution and larger decision space. Due to the time-frequency localization property of wavelet function, the learning characteristics are found be immensely improved as compared to the conventional DWNN [20]. These modified networks are also named as Wavelet Neural network (WNN)/WaveNets as they augment the learning potential of conventional neural network architecture with the identification and decomposition ability of wavelets.

Deep learning techniques have proved their superiority over traditional approaches for pattern recognition problems. However, the complexities associated with the overlapped and touching characters in a text requires even deeper approaches. This paper presents an intelligent and robust deep learning framework, DWNN for the text extraction from the images with overlapped and touching characters. The textual and image features distribution is very abruptly and randomly distributed in case of overlapped scripts, loosely configured characters, broken characters, connected characters. They are major cause of segmentation errors and result in inaccurate recognition. The application of high performance DWNN to learn and recognize the characters can greatly enhance the overall OCR performance. The major contributions of this research work can be mentioned as follows:

- Deriving the mathematical framework of the proposed DWNN using multiple layers.
- Exploiting the features distribution from even the local patches of the images through the localized spectral nature of activation function.
- Employing high dimensional deep feature representation.
- Analyzing the performance of the proposed framework for different challenging character variations.
- Attaining the best possible accuracy for the noisy, overlapped, and touching characters.

The paper is organized as follows: Section II deals with the literature survey through the analysis of related work. The mathematical framework of the proposed DWNN is given in Section III. The proposed DWNN based text extraction strategy for overlapped characters is discussed in Section IV. Effectiveness of the proposed strategy is illustrated through the simulation analysis in Section V while Section VI concludes the paper.

II. Related Work

The potential applications of OCR in various fields have attracted a lot of researchers to pursue research in this field. Numerous algorithms have been presented over the last decade for various stages of OCR viz. preprocessing, segmentation, feature extraction and recognition. Various techniques for the preprocessing of the images are presented by the researchers depending upon the type of images. Some commonly used techniques are noise removal, skew removal, thinning, morphological operations, etc. [21-23]. However, the most challenging aspect of OCR is the segmentation of images because of the diversity in the characteristics of text. It is also the most dominating phase of the OCR technique as the overall performance is depending upon the quality of segmentation. Because a single segmentation technique cannot be suitable for all type of textual distribution, many segmentation techniques have been proposed over the last decade [24].

Most of the segmentation techniques presented in the literature are based on the assumptions that the textual distribution is uniform, equidistant, and straight. But segmentation of overlapped and touching characters has not been addressed vastly and remained an open-ended problem. Farulla et al. [25] addressed this problem and proposed a fuzzy logic-based approach by combining various segmentation techniques altogether. Fuzzy rule base was prepared to derive a combined segmentation methodology which is robust to the noisy data. Garain and Chaudhuri [26] have implemented a multiple factors-based approach for the segmentation of touching characters in a printed text. The factors taken in the respective analysis were middleness, transitions and blob thickness. An algorithm was derived to facilitate the segmentation using these parameters. Nomura et al. [27] presented a histogram-based method for the primary character segmentation followed by morphological thickening and thinning operation to segment the overlapped characters. However, the accuracy was subject to the variation of textual data. A novel Harrow space filter was proposed by Tian et al.

[28] for the license plate character recognition. They have augmented weighted map algorithm to add robustness to the segmentation approach. Zheng et al. [29] have presented a segmentation technique for Arabic characters using the structural properties. They have utilized the vertical projections and some heuristics of these properties to differentiate between background and foreground regions to detect isolated characters. Similarly, the projections and statistical dimensional information was used for the Devanagari characters segmentation by Bansal and Sinha [30].

The problem of feature extraction and pattern recognition has also been addressed by the researchers to enhance the OCR performance. Various techniques like Syntactical Analysis, Neural Networks Template Matching, Hidden Markov Models, Bayesian Theory, etc. have been implemented for the efficient and robust recognition for different languages. E. B. Lacerda and C. A. Mello [31] presented a Self-Organizing Maps (SOM) based approach for the separating the touching characters. They have used the skeletonization process to cluster the feature points. Elnagar and Alhajj [32] proposed a technique to isolate the handwritten digit strings by normalizing and thinning which helped in identifying the feature points. These points are derived through the decision line from the deep points in the image. Gattal et al. [33] extended the research by combining different segmentation approaches based on configuration links between overlapped digits. They have used the sliding window Radon transform of these segmentation techniques to take the decision about selecting or discarding a digit image. Histograms of the vertical projection have also been used here for the contour analysis.

The computation cost associated with these segmentation techniques has posed serious concern during the real time implementation of OCR techniques. To deal with the issue of over-segmentation, several researchers have transformed the problem of segmentation and classification into a sole problem of recognition which does not involves these heavy segmentation algorithms. D. Ciresan [34] and A. G. Hochuli et al. [35] derived a convolutional neural network (CNN) framework for the recognition of these digits. These CNNs are trained over the datasets including isolated and touching text. Both these research works have avoided heavy segmentation issues, but their performance was subject to the availability of large datasets. H. Zhan et al. [36] have improved the OCR performance by combining Connectionist Temporal Classification (CTC) with Recurrent Neural Network (RNN). The features of the input image were extracted through the residual network and RNN was employed to derive the contextual information through these features. CTC model was derived to tune the parameters of the network for accurate classification. Zhang et al. [37] replaced RNN with DenseNet to further enhance the efficiency of the previous designs. These OCR frameworks have improved the recognition accuracy and efficiency, but the high computation complexity has remained the point of concern for the researchers.

Various configurations of deep learning have been applied for the problem of character recognition, but the performance of these techniques is subject to the availability of large training dataset. Also, their performance in dynamic and random environments cannot be guaranteed due to its dependency on initial conditions and uncertain convergence characteristics. The learning characteristics and weight adaptation is also found to be affected by the choice of a suitable activation function. Zhang et al. [38] has recently proposed a novel framework of neural network by replacing the sigmoid function by fast decaying wavelet function which presented better convergence rate and learning capabilities due to the space and frequency localization property of wavelets. Owing to the superior learning characteristics, this modified configuration of neural network known as Wavelet Neural network (WNN) has been used by the researchers in the fields of engineering where the feature space is highly dynamic and uncertain like object tracking, automatic control, advance communication, etc. [39]. WNN is used in this paper for the character recognition in overlapped and touching characters as the textual features distribution is uncertain and random in these types of images.

III. Deep Wavelet Neural Network Framework

The real potential of any neural network configuration is dependent on the activation function used for the learning of the network. Various activation functions like sigmoid, ReLU, fuzzy, etc. have been used by the researchers traditionally for variety of applications [40]. However, the performance of these functions for the data distribution with high diversity and randomness could not be guaranteed due to the globally defined nature. DWNN are the modified framework of conventional neural network architecture with translated and dilated versions of wavelet functions as activation function. The spectral characteristics of wavelets are explored to select the best suited wavelet for a typical data sample space. The dyadic translation and binary translation of the wavelets in a subspace $L^2(\Re)$ are used as basic functions for the data processing at the nodes of the wavelet network. Universal approximation property of neural network architecture deduces that the linearized combination of these basis functions is further used to evaluate the estimation function $\zeta \in L^2(\Re)$. A typical configuration of the DWNN is shown in Fig. 1. It is a four layered architecture namely input layer, wavelon layer, product layer and output layer. The preprocessing of the data is done at the input layer while the wavelon layer passes the data through the wavelet activation function. The deep features of the data are extracted at this layer through the translated and dilated versions on the wavelet function. The product of the outcome of these processed outcomes is then evaluated at the product layer and the decision is provided through the output layer.

The *n* dimensional biased network with *m* nodes generates the output as

$$\zeta = \omega^T \varphi(x,\tau,\sigma) + \mu^T \phi(x,\tau,\sigma) \tag{1}$$

where $x = [x_1, x_2, ..., x_n]^T \in R^n$ is the input vector, $\tau = [\tau_1, \tau_2, ..., \tau_m]^T \in R^{mxn}$ and $\sigma = [\sigma_1, \sigma_{2,}..., \sigma_m]^T \in R^{mxn}$ are translation and dilation parameters respectively, $\varphi = [\varphi_1, \varphi_{2,}..., \varphi_m]^T \in \Re^m$ is wavelet function and

$\phi = [\phi_1, \phi_2, ..., \phi_m]^T \in \Re^m$ represents the respective bias function. The weights and bias function of the network are represented as $\omega = [\omega_1, .., \omega_m]^T \in R^m$ and $\mu = [\mu_1, .., \mu_m]^T \in R^m$ respectively.

Optimization of the network parameters is the most important aspect of any deep learning network as it governs the nature of the learning curve of the network. The best approximation of the desired function can be attained through the optimal parameter vectors. The optimization of the network is achieved through an estimate function defined as:

$$\hat{\zeta}(x(n)) = \hat{\omega}^T \phi(x(n)) + \hat{\mu}^T \varphi(x(n)) \quad (2)$$

where $\hat{\omega}, \hat{\mu}$ are the estimates of the optimal values of the network parameters ω^*, μ^* respectively. The problem of optimization is reframed in terms of estimation error defined as

$$\tilde{\zeta}(x(n)) = \zeta(x(n)) - \hat{\zeta}(x(n)) = \left\{ \tilde{\omega}^T(n)\phi(x(n)) + \tilde{\mu}^T(n)\varphi(x(n)) + \varepsilon(x(n)) \right\} \quad (3)$$

The objective is now to reduce the value of estimation error $\tilde{\zeta}$ to an arbitrarily small value by carefully selecting the number of resolutions. The adaptive algorithm used to tune the weights of the NN framework is derived through gradient descent algorithm. The respective tuning laws for the weights are derived as:

$$\omega(n+1) = \omega(n) + \Delta\omega(n) = \omega(n) + \mu\left(-\frac{\partial e(n)}{\partial \omega(n)} \right) \quad (4)$$

where the learning rate and tuning weights are represented as μ and ω respectively. The weights are modified till the estimation error is not minimized to exceedingly small value.

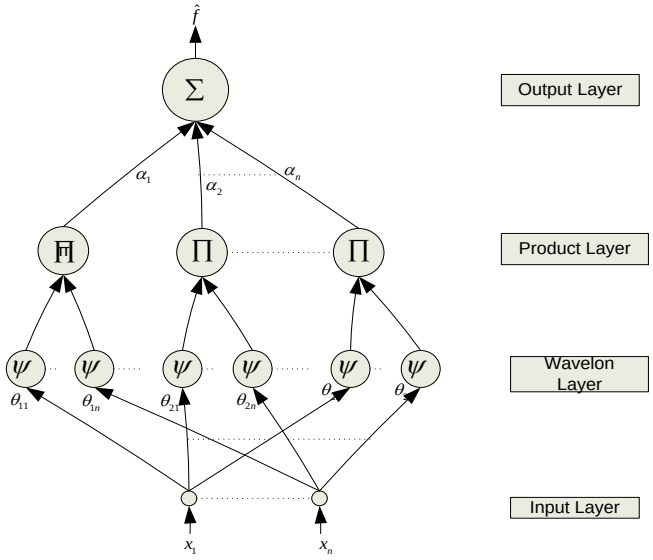


Fig. 1. Architecture of Wavelet Neural Network (WNN).

IV. Proposed Methodology

The text recognition framework for the overlapped characters using DWNN is shown in Fig. 2. The overall process is divided into following phases: Preprocessing, segmentation, Classification, and recognition. The proposed framework for the DWNN based OCR for overlapped characters is shown in Fig. 2.The process is discussed in detail below:

Preprocessing: The preliminary process icing of the images is performed at preprocessing stage which includes binarization, noise filtering, skew correction and thinning. Binarization is performed over the image to transform it from RGB to grayscale. A threshold is chosen in this process to separates the foreground and background information. However, it is overly complex problem in cases where the contrast between text pixels and background is low. The thin text strokes or non-uniform illumination during image capturing may result in background bleeding into text pixels during digitization. Multi-thresholding is performed here to identify the relevant gray level information and eliminate the noise and isolated points. The pixel values of the input image are first provided to the multi-threshold module which compares their values with the upper and lower threshold values. If the data is within the range defined through the threshold levels, the pixel data is assigned as 1, otherwise it will be assigned the value of 0. It results into the binary distribution of the text image which could further be used for the respective analysis. The resultant digital image may be subject to some disturbances or noise due to the undesired aspects in optical scanning devices and camera. This noise is removed during preprocessing through suitable filters. The typical noise present in the scanning process is salt and pepper noise because of the quality of paper scanned and parasitic components. The filtering of these noises may result into some edge losses; therefore, median filter is used in this work so as to preserve the edges of the text images.

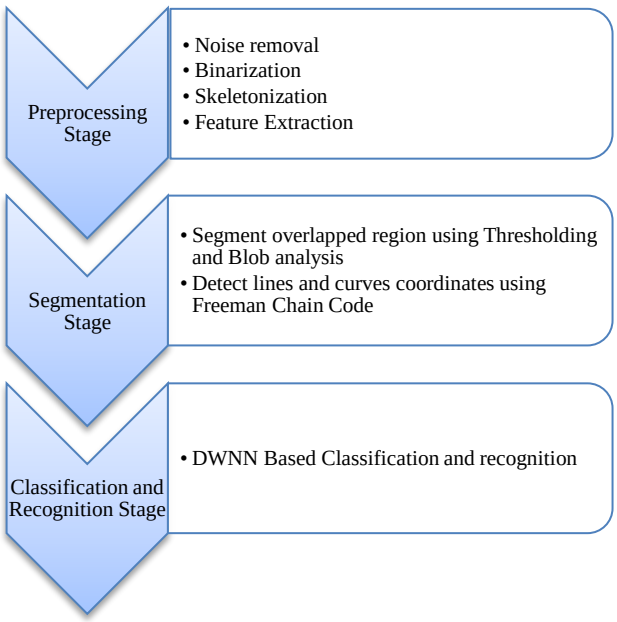


Fig. 2. Proposed Methodology.

Skew correction is then performed over the filtered images to remove any kind of disorientation of text in the images. The skew of the scanned document image specifies the deviation of its text lines from the horizontal or vertical axis. The projection profiles are used as a suitable feature for skew detection after extracting the black text pixels for analysis. The shape of the characters is then extracted using the thinning process by exploring the pixel distribution of the characters smoothly. The process of preprocessing used in this work is shown in Fig. 3.

Segmentation: The most important process of the proposed work is segmentation which is responsible for efficient split of the overlapping and touching characters in a text image. The overall performance of the OCR is dependent on the correctness of segmentation as the corresponding features of the segmented images are used for the recognition and classification. Keeping the seriousness of this process in view, Blob analysis-based method is used for the segmentation in this paper. The filtered pixels are classified or clustered in Blob analysis based on the respective pixel values. If the pixel value is nearly equal to the neighboring pixel value, then they are kept in a same blob. This process results into number of clusters or blobs having same kind of pixel distribution and spectral characteristics. The segmentation in this paper is performed using the blob adjacency analysis where each pixel is checked with its eight neighbor pixels on vertical, horizontal, and diagonal axes.

The projection profiles of these blobs and the connected component analysis are used to derive the blob classification. The performance for the overlapped and touching characters is further improved by applying the dilation and erosion over the merged characters.

The shape and size of the blobs are identified in this work using the Freeman chain coding with eight connectivity approach which derives the boundary of the blob contour. Chain codes use the connected sequence of straight-line segments of specified length and orientation to identify the boundary. Freeman chain coding is a linear structure derived through the quantization of the trajectory traced by the centers of adjacent boundary elements in a pixel array. The respective code is generated by the clockwise or anticlockwise scanning of the boundary and assigning the orientation value to the segment connecting each pair of pixels. This process for the shape and size identification results into a compact and translation independent representation of a binary contour. It also provides a lossless compression and preserving capacity for all morphological or topological information.

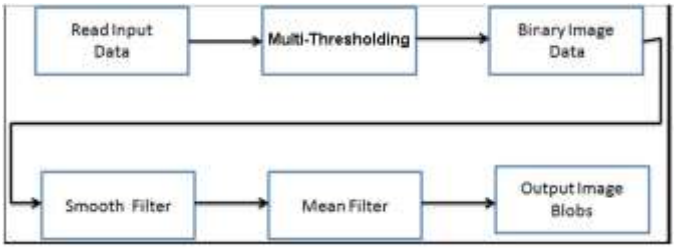


Fig. 3. Preprocessing Framework.

Classification and Recognition: Various features of blobs are extracted in this work to derive the dataset which is used for the training of the deep learning network. The selection of the features plays a particularly important role in defining the performance of the character recognition. In this work, the parameters selected and extracted for the training are based upon the blob analysis. These features resemble the shape, texture, blob area, perimeter, corners, etc. which are generated through the skeletonization process. The respective feature vector is derived by dividing the blob zones into various windows of equal size and the respective chain codes are used to derive the parameters like number of horizontal lines, vertical lines, right diagonal lines and left diagonal lines. These features are then used for the training of the proposed DWNN framework. It is responsible for the final decision making about the characters which are to be extracted from the text. The feature dataset is divided into two parts, training dataset, and testing dataset. The derived feature vector is denoted as X as.

$$X = f_1, f_2, \ldots, f_n \tag{5}$$

Where n denotes the number of blob zones and f represents the respective feature sets. DWNN is trained with these feature sets to derive a recognition model for the overlapped and touching characters in the text.

The recognition model is derived from (1) as

$$\zeta = \omega^T \varphi\left(X, \tau, \sigma\right) + \mu^T \phi\left(X, \tau, \sigma\right) \tag{6}$$

The wavelet (φ) used in the proposed DWNN model is Mexican hat wavelet which has shown best resolution performance in analyzing the sharp features in the data distribution. It has two side-lobes and central peak as shown in Fig. 4 and therefore known as second Gaussian wavelet (g_2 wavelet) as well. It is the negative normalized second derivative of a Gaussian function. The mathematical description of Mexican hat wavelet can be derived by evaluating the second derivative of the Gaussian probability density function (pdf) as

$$\varphi(t) = \frac{2}{\pi^{\frac{1}{4}}\sqrt{3\rho}}\left(1 - \frac{t^2}{\rho^2}\right)e^{\frac{-t^2}{2\rho^2}} \tag{7}$$

where ρ represents the standard deviation of the Gaussian pdf.

The proposed DWNN model used for the text extraction comprises of three nodes and the weights and bias of the network are tuned and optimized using the gradient descent algorithm over the error value derived in (4) as:

$$\omega(n+1) = \omega(n) + \Delta\omega(n) = \omega(n) + \mu\left(-\frac{\partial e(n)}{\partial \omega(n)}\right) \tag{8}$$

The recognition accuracy is considered to evaluate the performance of the proposed DWNN based text extraction strategy.

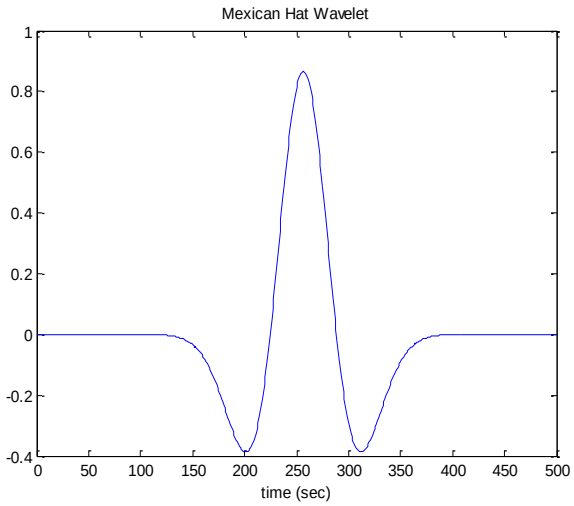


Fig. 4. Mexican Hat Wavelet.

V. Experiment Results and Discussion

The performance of the proposed DWNN based text recognition with overlapped characters is assessed through the experimental analysis in MATLAB. The recognition capability of the proposed deep learning network is evaluated based on some metrics namely estimation error, cost function and Accuracy. More than 100 images with overlapped and non-overlapped characters are taken for the experiment with 80 % used for training and 20% for testing purpose. Fig. 5 shows the variation of cost function with respect to time and reflects the efficacy of the optimization algorithm for the tuning of the proposed DWNN. The variation of estimation error is shown in Fig. 6 which clearly reflects the learning characteristics of the network. The values of the estimation error are converging and ranging around zero as the learning goes on.

The performance of the proposed DWNN based character recognition is also shown by the accurate segmentation of the characters in the input images. The input images along with their respective segmentations are shown in Fig. 7 to 10. Although the accuracy of OCR in case of overlapped or touching character solely depends upon the quality of segmentation which relies on the quality of the image, the recognition capability has been evaluated in terms of average accuracy and found to be around 94 percent. The fast convergence rate of the proposed deep learning network can easily been deduced from the sharp decrease in the value of estimation error. It also represents the impact of using wavelet function as the activation function in the conventional neural network framework. The capability of wavelet as kernel in the neural network framework is also reflected from the efficient classification.

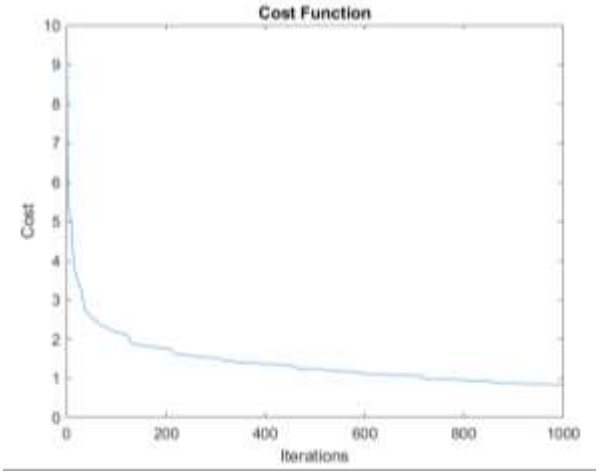


Fig. 5. Cost Function v/s Time.

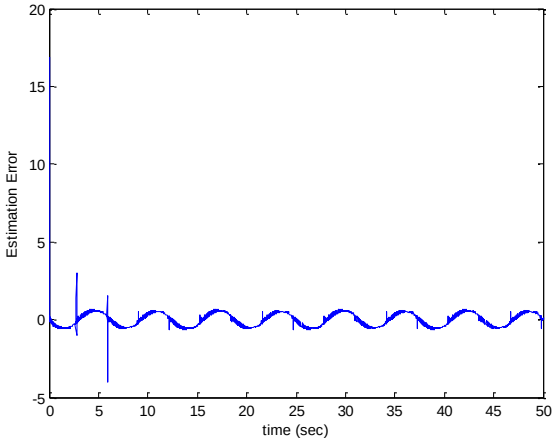


Fig. 6. Estimation Error v/s Time.

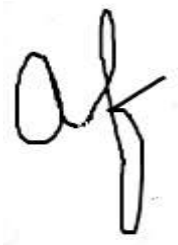


Fig. 7. Input Image 1.

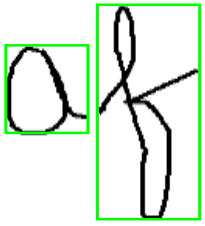


Fig. 8. Segmented Image 1.

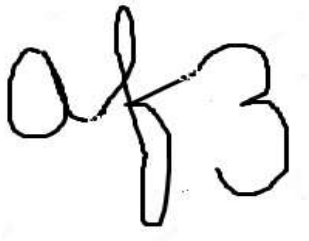


Fig. 9. Input Image 2.

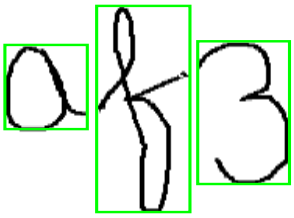


Fig. 10. Segmented Image 2.

VI. Conclusion

A deep learning based intelligent text recognition is presented in the paper which has used DWNN to enhance the learning capabilities of the conventional deep learning models. Mexican hat wavelet has been used as the activation function to mitigate the impact of uncertainties and randomness in the data distribution over an image. The challenging problem of segmentation and recognition of overlapped or touching characters has been addressed in this paper. The segmentation is achieved through the combination of blob adjacency analysis and Freeman chain coding technique. The extracted features are used for the training of the proposed DWNN model. Cascaded layered architecture of the translated and dilated versions of wavelet function is used to derive the complete framework of the DWNN. Tuning laws have been derived through gradient descent algorithm to achieve the optimal learning characteristics. The performance of the proposed deep learning framework is evaluated through the experimental analysis in terms of estimation error and cost function which has reached to a desirable range within an exceedingly small amount of time. The segmentation part of the process is found to be the most crucial aspect of OCR and it is subject to the nature of connectivity between the characters. The future aspect of this research is to add more robustness during the segmentation process to make the proposed model more effective even for the characters with high degree of superposition and overlapping.

References

[1] S. Nomura, K. Yamanaka, O. Katai, H. Kawakami, and T. Shiose, "A novel adaptive morphological approach for degraded character image segmentation,"Pattern Recognit., vol. 38, no. 11, pp. 1961-1975, Nov. 2005.

[2] A. Rehman and T. Saba, "Performance analysis of character segmentation approach for cursive script recognition on benchmark database," Digit.Signal Process., vol. 21, no. 3, pp. 486-490, May 2011.

[3] P. P. Roy, U. Pal, J. Lladós, and M. Delalandre, "Multi-oriented touching text character segmentation in graphical documents using dynamic programming,"Pattern Recognit., vol. 45, no. 5, pp. 1972-1983, May 2012.

[4] N. Arica and F. T.Yarman-Vural, "Optical character recognition for cursive handwriting," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 6,pp. 801-813, Jun. 2002.

[5] K. S. Dash, N. B. Puhan, and G. Panda, "BESAC: Binary external symmetry axis constellation for unconstrained handwritten character recognition,"Pattern Recognit. Lett., vol. 83, no. 3, pp. 413-422, Nov. 2016.

[6] K. S. Dash, N. B. Puhan, and G. Panda, "Handwritten numeral recognition using non-redundant Stockwell transform and bio-inspired optimal zoning," IET Image Process., vol. 9, no. 10, pp. 874-882, Sep. 2015.

[7] G. Raju, B. S. Moni, and M. S. Nair, "A novel handwritten character recognition system using gradient based features and run length count," Sadhana, vol. 39, no. 6, pp. 1333-1355, Dec. 2014.

[8] P. Singh, A. Verma, and N. S. Chaudhari, "Feature selection based classifier combination approach for handwritten Devanagari numeral recognition,"Sadhana, vol. 40, no. 6, pp. 1701-1714, Sep. 2015.

[9] U. Garain and B. B. Chaudhuri, "Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multi-factorial analysis," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 32, no. 4, pp. 449-459, Nov. 2002.

[10] Y.-K. Chen and J.-F.Wang, "Segmentation of single- or multiple-touching handwritten numeral string using background and foreground analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 11, pp. 1304-1317,Nov. 2000.

[11] V. Bansal and R. M. K. Sinha, "Segmentation of touching and fused Devanagari characters," Pattern Recognit., vol. 35, no. 4, pp. 875-893,Apr. 2002.

[12] P. Sahare and S. B. Dhok, "Review of text extraction algorithms for scene text and document images," IETE Tech. Rev., vol. 34, no. 2, pp. 144-164,Apr. 2016.

[13] B. Zhu and M. Nakagawa, "A robust method for coarse classifier construction from a large number of basic recognizers for on-line handwritten Chinese/Japanese character recognition," Pattern Recognit., vol. 47, no. 2,pp. 685-693, Feb. 2014.

[14] G. A. Montazer, H. Q. Saremi, and V. Khatibi, "A neuro-fuzzy inference engine for Farsi numeral characters recognition," Expert Syst. Appl.,vol. 37, no. 9, pp. 6327-6337, Sep. 2010.

[15] K. Verma and R. K. Sharma, "Comparison of HMM and SVM-based stroke classifiers for Gurmukhi script," Neural Comput. Appl., vol. 28, pp. 51-63, Dec. 2016.

[16] A. A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network," Pattern Recognit., vol. 43, no. 7,pp. 2582-2589, Jul. 2010.

[17] Sampath AK, Gomathi N,"Decision tree and deep learning based probabilistic model for characterrecognition", J Cent S Univ 24:2862–2876, 2017.

[18] Semwal VB, Gaud N, Nandi GC, "Human gait state prediction using cellular automata and classificationusing ELM", In: Machine Intelligence and Signal Analysis, p 135-145, 2019.

[19] M. Sharma, A. Kulkarni and A.Verma, "Wavelet Adaptive Output Tracking Control for a Class of Delayed Uncertain MIMO Nonlinear Systems Subjected to Actuator Saturation", Proceedings of IEEE international conference on advances in computing, control and telecommunication technologies,ACT-09, India, pp. 705-710, 2009.

[20] M. Sharma, A. Kulkarni and A.Verma, "Wavelet adaptive observer based control for a class of uncertain time delay nonlinear systems with input constraints", Proceedings of IEEE international conference on advances in recent technologies in communication and computing,ARTCOM-09, India, pp. 863-867, 2009.

[21] Guruprasad P, Majumdar J,"Multimodal recognition framework: an accurate and powerful Nandinagari handwritten character recognition",Modelb. Procedia Comput Sci 89:836–844, 2016.

[22] Kanimozhi VM, Muthumani I,"Optical character recognition for English and Tamil script",International Journal of Computer Science and Information Technologies 5(2):1008–1010,2014.

[23] Priyadarshni, Sohal JS,"Improvement of artificial neural network based character recognition system",using SciLab. 10:1-9,2016.

[24] A. Antonacopoulos and D. Karatzas, "Semantics-Based Content Extraction in Typewritten Historical Documents," in Eighth International Conference on Document Analysis and Recognition, pp. 48-53, 2005.

[25] G. A. Farulla, N. Murru, and R. Rossini, “A fuzzy approach to segment touching characters,” Expert Syst. Appl., vol. 88, pp. 1-13, Dec. 2017.

[26] U. Garain and B. B. Chaudhuri, “Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multi-factorial analysis,”IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 32, no. 4, pp. 449-459,Nov. 2002.

[27] S. Nomura, K. Yamanaka, O. Katai, H. Kawakami, and T. Shiose, “A novel adaptive morphological approach for degraded character image segmentation,”Pattern Recognit., vol. 38, no. 11, pp. 1961-1975, Nov. 2005.

[28] J. Tian, R. Wang, G. Wang, J. Liu, and Y. Xia, “A two-stage character segmentation method for Chinese license plate,”Comput. Elect. Eng., vol. 46, pp. 539-553, Aug. 2015.

[29] L. Zheng, A. H. Hassin, and X. Tang, “A new algorithm for machineprinted Arabic character segmentation,” Pattern. Recogn. Lett., vol. 25,no. 15, pp. 1723-1729, Nov. 2004.

[30] V. Bansal and R. M. K. Sinha, “Segmentation of touching and fused Devanagari characters,” Pattern Recognit., vol. 35, no. 4, pp. 875-893,Apr. 2002.

[31] E. B. Lacerda and C. A. Mello, “Segmentation of connected handwritten digits using self-organizing maps,” Expert Syst. Appl., vol. 40, no. 15,pp. 5867-5877, Nov. 2013.

[32] A. Elnagar and R. Alhajj, “Segmentation of connected handwrittennumeral strings,” Pattern Recognit., vol. 36, no. 3, pp. 625-634, Mar. 2003.

[33] A. Gattal and Y. Chibani, “SVM-based segmentation-verification of handwritten connected digits using the oriented sliding window,” Int. J. Comput. Intell. Appl., vol. 14, no. 1, Mar. 2015.

[34] D. Ciresan, “Avoiding segmentation in multi-digit numeral string recognition by combining single and two-digit classifiers trained without negative examples,” in Proc. 10th Int. Symp. Symbolic Numeric Algorithms Sci.Comput., Sep. 2008, pp. 225-230.

[35] A. G. Hochuli, L. S. Oliveira, A. Britto, Jr., and R. Sabourin, “Handwritten digit segmentation: Is it still necessary?” Pattern Recognit., vol. 78,pp. 1-11, Jun. 2018.

[36] H. Zhan, S. Lyu, and Y. Lu, “Handwritten digit string recognition using convolutional neural network,” in Proc. 24th Int. Conf. Pattern Recognit. (ICPR), Sep. 2018, pp. 3729-3734.

[37] K. Zhang, Z. Zhang, Z. Li, andY. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process.Lett., vol. 23, no. 10, pp. 1499-1503, Oct. 2016.

[38] Q. Zhang and A. Benveniste, “Wavelet networks,” IEEE Transactions on Neural Networks, Vol. 3, no. 6, pp.889-898, November 1992.

[39] J. Zhang, G. G. Walter, Y. Miao, and. W. Lee, “Wavelet neural networks for function learning,” IEEE Transactions on Signal Processing, Vol. 43, no. 6, pp.1485-1497, June 1995.

[40] B.Delyon, A. Juditsky, and A. Benveniste, “Accuracy analysis for wavelet approximations,” IEEE Transactions on Neural Networks, Vol. 6, no. 2, pp.332-348 March 1995.

Performance Improvement of Network Coding for Heterogeneous Data Items with Scheduling Algorithms in Wireless Broadcast

Romana Rahman Ema[1], Md. Alam Hossain[2]*, Nazmul Hossain[3], Syed Md. Galib[4], Md. Shafiuzzaman[5]
Computer Science and Engineering Department, Jashore University of Science and Technology, Jashore-7408, Bangladesh

***Abstract*—This is the age of information. Now-a-days everyone communicates with each other by means of digital systems. Humans are always communicating with each other on the go. On-demand broadcasting is an efficient way to broadcast information according to user requests. In an on-demand broadcasting network, anyone can satisfy multiple clients in one broadcast which helps to fulfill the enormous demand of information by clients. The optimized flow of digital data in a network through the transmission of digital evidence about messages is called network coding. The "digital evidence" is composed of two or more messages. Network coding incorporated with data scheduling algorithms can further improve the performance of on-demand broadcasting networks. Using network coding, anyone can broadcast multiple data items using single broadcast strategy which can satisfy the needs of more clients. In this work, it is described that network coding cannot always maintain its superiority over non-network coding when the system handles different sized data items. However, the causes of performance reduction on network coding have been analyzed and THETA based dynamic threshold value integration strategy has been proposed through which the network coding can overcome its limitation for handling heterogeneous data items. In the proposed strategy, THETA based dynamic threshold will control which data item will be selected from the Client Relationship graph (CR-graph) so that large sized data items cannot be encoded with small sized data items. Simulation result shows some interesting performance comparison.**

***Keywords*—Network coding; scheduling algorithms; CR graph; wireless broadcast; simulation; LTSF; STOBS; performance metric**

I. Introduction

Now-a-days almost everyone carries a portable cellular computing device from a laptop computer to smartphone. All these devices share information to the network on the go. This also requires an infrastructure that does not require a user to maintain a fixed connection in the network and allows mobility. Wireless networks require mobility, distributed sensing and city-wide internet connectivity. For broadcasting the data to the client, network coding uses the limited bandwidth of the wireless efficiently [17] [23]. Network Coding, as a field of study is young which was first introduced in [27] [30]. It is a new concept. Study on the performance of network coding shows that it can utilize the available limited bandwidth of the network to achieve improved throughput in multicast communication [16] [31] [32]. Network coding is applied on on-demand broadcasting network [14] [23] [28]. Here the server broadcasting the data has the information of every client it is broadcasting. Server uses this information to keep track of the data received by clients. Then the server encodes data and broadcasts them on the network. All the clients receive the encoded data and use its own received data to decode the encoded data. Using network coding, a server can serve multiple requests at the same time [17] [22].

Network coding can increase the performance of a broadcasting network in many aspects. It increases throughput, robustness, security in network as well as decreases deadline miss ratio, stretch, response time [17]. But while working on heterogeneous data items, network coding has some drawbacks [9]. It does not perform well as it has been on singular data items [29]. It is caused by the encoding technique which is used in network coding [15]. In XOR encoding, we encode the data items that are found in the maximum clique from the CR-graph [4] [7] [24] [25]. CR-graph is constructed through the data regarding clients' relationships of requested and cached data items [4] [7]. In the proposed THETA based dynamic threshold value integration strategy, the drawbacks of the traditional network coding approach in the scenario of heterogeneous data items have been minimized. Large sized data items and small sized data items have been filtered and encoded separately for improving the performance of network coding.

The rest of this paper is organized as follows. Section II contains related work. Section III illustrates the system model for implementing our proposed strategy. Section IV describes the performance evaluation. Our final thoughts are included in Section V.

II. Related Work

G. G. Md. Nawaz Ali, Yuxuan Meng et al. [1] performed simulation-based analysis based on top of generalized encoding model on both in homogenous as well as the heterogeneous environment for measuring the effectiveness plus adaptability of network coding assisted scheduling algorithms. They analyzed the performance of diverse scheduling algorithms both in non-coding then their proposed coding method utilizing dissimilar performance metrics.

Yuxuan Meng, Edward Chan at al. [2] analyzed the effect of network coding with different scheduling algorithms. They conducted various experiments to measure the performance of broadcasts considering standard access moment, due date ignores relative amount along with typical stretch out.

*Corresponding Author

Cheng Zhan, Victor C. S. Lee et al. [3] proposed a generalized framework so that data scheduling algorithms can be incorporated with network coding for broadcasting on demand requests. They described that with coding, performance can be improved using different scheduling algorithms.

Jun Chen, Victor C. S. Lee et al. [4] proposed a new coding strategy named AC, for implementing an efficient coding mechanism. They also proposed two coding assisted algorithms named ADC-1 and ADC-2 considering data scheduling and network coding. Their simulation results showed that response time was dynamically reduced using both ADC-1 and ADC-2. They also showed that ADC-1 and ADC-2 performed better than conventional and other coding assisted algorithms.

Mohamed A. Sharaf and Panos K. Chrysanthis [8] proposed a new scheduling algorithm named STOBS-α for grouping requests and only one-time delivery of broadcasting results to the clients. Their proposed heuristic on demand algorithm was experimented using access time and fairness for mobile clients.

III. Need of the Improvements

From studies it is noted that when there is no difference in data item size, there is no problem in encoding. For instance, if it is needed to encode three data item d1, d2 and d3 of unit size 1, the size of encoded data item d1⊕d2⊕d3 is also 1. But when we have to encode data items with different size then there is a slight problem. In this condition, the encoded data item's size is the size of the largest item selected for encoding. Let the size of d1 is 1 unit, d2 is 3 unit and d3 is 7 unit. Then the size of encoded data item d1⊕d2⊕d3 is 7 unit. In traditional network coding, large data items are selected with small data items for encoding which in terms cause performance reduction. That also leads to increased stretch and response time, thus hampering the performance of the network [12]. Traditional scheduling algorithms [5] [12] [18] are able to perform better than network coding in such conditions. For this reason, a new modified strategy in network coding has been established to handle heterogeneous data items with ease for maintaining an improved throughput, stretch and response time. The contribution of this paper is as follows:

1) To design a system model, where the server maintains the specification of network coding.

2) To implement the proposed modified strategy which will eliminate the drawback of network coding for heterogeneous data items.

3) To simulate, integrate and analyze our proposed approach with other existing basic scheduling algorithms and compare their performances.

IV. System Model

A. System Architecture

To fulfill more requests earlier than their due dates as well as to assure operative utilization of the constrained bandwidth are the main goals of real-time scheduling and coding. Our system architecture is based on top of the conventional on demand broadcast set-up [4] [7] [10] [14] [18]. The architecture is shown in Fig. 1. The system is set aside by one server with a number of end devices. All end devices have a local cache along with provisions for a certain data core which is broadcasted by the server [1] [13]. Due in the direction of the obligatory room of the end device's caches, a certain guiding principle is applied intended for cache substitution. If the inquired data core cannot be initiated in its cache, the end device sends a request, and its active cache stand-in data to the server through an uplink tunnel [1]. All requests conceivably will necessitate auxiliary data portion from the server. Later than transfer requests headed for the server, end devices listen to the broadcast tunnel to recapture their requested data [1] [13]. It is presupposing that an end device doesn't cache this arriving encoded information but it cannot decode any asked data piece by utilizing this encoded information. If an end device gets and decodes every requested data substance earlier than the time limit, in that case, the requests can be content. In other cases, the request misses its time limit as well as there is no value to the end device [4].

On receiving a request, the server embeds it into a service queue. A request holds up to be scheduled in the service queue until every one of its requested data substances are broadcasted otherwise it gets to be infeasible for scheduling [1] [20]. When the leftover slow-moving phase is smaller than the compulsory phase obligatory towards broadcast every one of the leftover unprocessed data substances, the appeal is considered impossible to be scheduled [1]. A request is removed from the service queue and becomes infeasible, if it misses its required deadline [1]. The server primarily recovers the asked information substance put away within the local database based on top of certain scheduling algorithms then, in that case, encodes the data substance based on data concerning end devices' cached and requested data substance. Lastly, the server broadcasts the encoded information via the downlink tunnel. Inside our model, server and end devices purely exploit the basic XOR operations to encode and decode information [3] [7] [30]. Therefore, the encoding, as well as decoding operating cost and hold-up, can be overlooked.

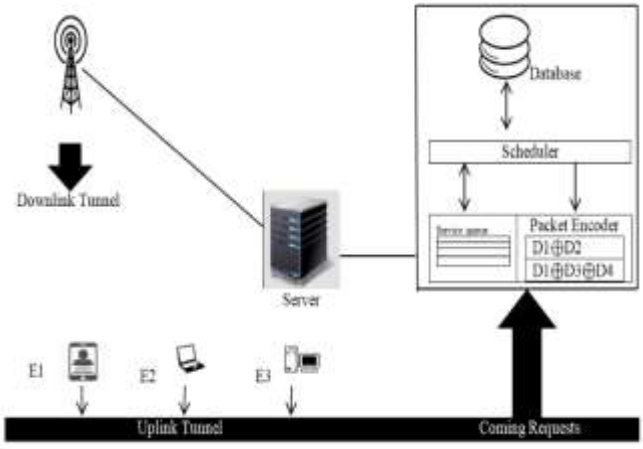


Fig. 1. System Architecture.

B. Graph Model

Our graph model is based on the graph model proposed and discussed by Zhan at al. [3]. In this approach, the CR-graph is constructed on the THETA based threshold mechanism basis. The system has a data server S and n number of end devices E = {e_1, e_2, ... e_n}. Set X(e_i) denotes the set of requested data items of end devices' e_i and set Y(e_i) denotes the set of cached data items of end devices' e_i. The server has a database which contains all the data items. Set m denotes the overall data items contained in the database D. Definition 1: Given E = e_1, e_2, ... e_n, D = d_1, d_2,.... d_m, X(e_i) ⊆ D, Y(e_i) ⊆ D, X(e_i) ∩ Y(e_i) = ∅ [3][6]. A graph G(V, E) can be built the same as follows:

- V= { v_{ij} | end device e_i requests for data item d_j, 1≤ i≤ n, 1≤j≤m}
- E = (v_{i1j1}, v_{i2j2}) | j_1=j_2; or $j_1 \neq j_2$, $d_{j2} \subseteq Y(u_{i1})$, $d_{j1} \subseteq Y(u_{i2})$, | SizeOF($d_1$)- SizeOF($d_2$) | THRESHOLD

If we weigh up on-demand broadcast circumstances in Fig. 2(a) which consists of a server, S and five end devices, e_1, e_2, e_3, e_4 and e_5. Presume that the server has four data substances d_1, d_2, d_3 and d_4. If we assume that end device e_1 has already stored d_2, d_3, d_4 in its cache from preceding broadcasting, end device e_2 has d_1, d_2, d_4 in its cache, end device e_3 has d_2, d_3, d_4 in its cache, end device e_4 has d_1, d_2, d_4 in its cache and end device e_5 has d_1, d_2, d_3 in its cache. Now if we assume that end device e_1 is requesting data item d_1, end device e_2 and e_3 are requesting data item d_2, end device e_4 is requesting data item d_3 and end device e_5 is requesting data item d_4. The data sizes of d_1, d_2, d_3 and d_4 are 1 unit in addition to the broadcast transmission capacity is B=1, which infers the server can broadcast one information piece every time unit.

As of explanation 1, the diagram matching to Fig. 2 is developed as Fig. 3. Within this stature vertex V_{11} speaks to the request from end device e_1 for data d_1. End device e_1 has d_2 requested by e_2 and end device e_2 has d_1 requested by e_1, there is an edge (V_{11}, V_{22}). It is also shown that the end device e_3 has d_3 requested by e_4 and the end device e_4 has d_2 requested by e_3, there is an edge (V_{32}, V_{43}). Other edges are constructed by following the same rule.

Requested Data Item	d_1	d_2	d_2	d_3	d_4
End devices	e_1	e_2	e_3	e_4	e_5
Cached Data Item	d_2,d_3,d_4	d_1,d_2,d_4	d_2,d_3,d_4	d_1,d_2,d_4	d_1,d_2,d_3

Fig. 2. A Demo of On-demand Information Broadcast.

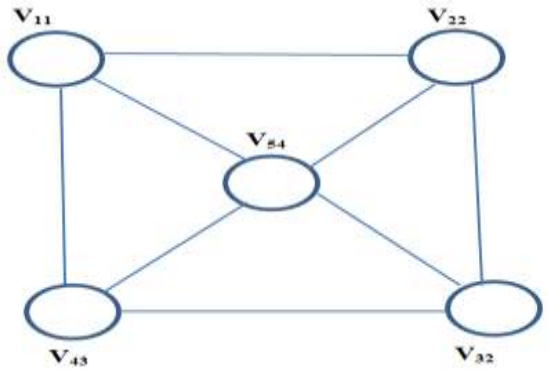


Fig. 3. CR-Graph Developed Commencing the Case in Fig. 2.

C. Proposed THETA based Dynamic Threshold Calculation Strategy

The proposed THETA based dynamic threshold integration is based on the fact that large data sized items will not be encoded with small sized data items.

We construct the undirected graph G (V, E) according to G (V, E) as mentioned in III (B) section. Candidate vertex v_{mi} (denotes data item d_i requested by end device E_m) need to be initiated with V(G).

- For each $v_{ij} \in V(G)$ do
- if SizeOf (d_k) < 100 then
- THRESHOLD= (ln (SizeOf(d_k)))^2
- end if
- if SizeOf (d_k) > 100 then
- THRESHOLD= (SizeOf(d_k) ÷ ln (SizeOf(d_k)))
- End

We need to make a decision which data items would be encoded together. For this reason, we have used dynamic thresholds. This gives better results. Threshold calculating process is given above. Here we have given an example. Suppose we need to broadcast 5-unit data items and 50-unit data items. If we choose 5-unit data items as candidates, the threshold would be around 3. So, 50 unit sized data items won't be in a set with a 5-unit data item set. Only, data item size in between 2 and 8 would come to consideration. Again, when 50-unit data items are candidate, the threshold would be around 15. So, 5 unit sized data item won't be in a set with 50. In this case the range of encoding would be 35 to 65. Though they produce different thresholds, both will face less stretch.

The proposed approach is to change THETA for each candidate by doing THETA= candidate requests requested data item size ÷ 3. But it is tried to choose THETA through a general equation. The proposed algorithm strategy is given below.

- At first, we make pairs which contain Client_Id and Requested_Data_Id.
- Then, we choose each request as a candidate request one by one. For calculating dynamic THETA, we have

used candidate requests requested data item size. Here, THETA is equal to either square of ln (candidate requests requested data item size) or candidate requests requested data item size ÷ ln (candidate requests requested data item size).

- By using THETA, we make a set. If the difference of candidate request's requested data item size and other requests requested data item size is less than THETA then this request is inserted into the set. We repeat this process for all candidate requests.
- From these sets, we find out maximum clique and then do network coding by broadcasting data.

If two vertices of a similar subset are linked through an edge in an undirected graph, it is considered as a clique [4] [26]. There are a number of preferences in the proposed methodology. We are using THETA based dynamic threshold integration. It helps to separate small data items from large data items. If small data items are encoded with large data items, they face large stretch and response time also increases.

V. Performance Evaluation

A. Overview of Comparable Scheduling Algorithm

With the rapid growth of on-demand broadcasting networks, servers have to serve significantly large numbers of clients every day and it is always increasing. So to balance out the increasing load of servers, the necessity of new and improved scheduling algorithms is very high. In network coding, we have to incorporate scheduling algorithms according to our framework so that they maintain their characteristics for scheduling data items for normal networks. Two algorithms have been implemented using the system model (III-A). For heterogeneous data items, STOBS and LTSF scheduling algorithms perform efficiently better than other scheduling algorithms (FCFS, MRF, LMF and others) as those algorithms are generally implemented for single data item [11].

1) STOBS (Summary Tables On-Demand Broadcast Scheduler): In STOBS, the server maintains a queue to store the requests of clients at the time of their arrival. This scheduler chooses a data item for broadcasting with highest (R*W)/S [3] [8] [12] [19]. A summary table T^x is maintained for each request of Q^x. The server keeps the following information [8] [12] [19].

- R: Number of arrival requests for the table $T^{x.}$ When a request for T^x arrives, the value of R increments.
- W: Waiting time of the first request $Q^{x.}$
- S: Table size.

2) LTSF (Longest Total Stretch First): LTSF chooses a data item intended for broadcasting concurring to the order of the maximum total recent stretch [3] [12]. The data piece having the utmost whole current distend is broadcasted earliest [18]. Recent stretch records are calculated by the ratio of the waiting time of pending requests to its time of service [3] [9] [12] [18] [21].

B. Performance Metrics

1) Average response time: Average response time is the ratio of the summation of altogether request's response time in the direction of the entire number of requests [8] [18] [21]. Requests are served quickly if the value of average response time is low.

2) Average stretch: Minimizing the average stretch for heterogeneous data items is the main issue considered for scheduling algorithms. We find average stretch in our simulation model using the following equation.

Average Stretch=Total response time for all end devices/Total service time for all end devices [8] [18] [21].

C. Simulation Model

We performed detailed analysis using CSIM19 [11]. The simulation parameters used for our system architecture (III-A) is shown in Table I. Most of these parameter values are considered from related works [1] [2] [3] [4] [5] [7]. In our simulation model, the server maintains a cache for every end device. At first end devices are generated with data items in their cache automatically. Then, they make their requests to the server maintaining an inter arrival time (IATM). We use IATM in accordance with average data item size. In our simulation:

IATM = 100/Average data item size;

If IATM is low, then end devices will make their requests more frequently which can overload the server.

TABLE I. Simulation Parameters

Parameters	Default	Range	Description
IATM	-	-	Request generation interval
NUMENDDEVICE	100	25-600	Number of end devices
DBSIZE	1000	-	Quantity of information objects in the database
BANDWIDTH	1KB/sec	-	Broadcasting bandwidth
THETA (θ)	0.4	0.2-1.0	Zipf distribution parameter.
CACHSIZE	10	30-180	The maximum amount of data stored in every client's cache.
DWNSIZEMIN	1Kb	-	Minimum size of data item in database
DWNSIZEMAX	60Kb	30-70	Maximum size of data item in database

D. Performance Analysis

The proposed THETA based dynamic threshold integration has been implemented using the system model described in III-A. Overall performance is analyzed and compared using two metrics: average response time and average stretch. We measured the performance by varying item size and cache size with our dynamically changing THETA. Simulation results show that there is significant increase of performance in the network coding environment with our proposed THETA based dynamic threshold

calculation strategy. The reason behind performance improvement is large sized data items are not being selected with small sized data items with our proposed strategy.

1) Performance comparison of average response time for varying item size with different cache size: Tables II, III and Fig. 4(a), 4(b) shows the observations of average response time by varying item size for different cache size. For both STOBS and LTSF algorithms, there is significant increase of performance in the network coding environment. For STOBS, average response time decreases by .25 percent for cache size 30 and .095 percent for cache size 60. For LTSF, average response time decreases by .25 percent for cache size 30 and .26 percent for cache size 60.

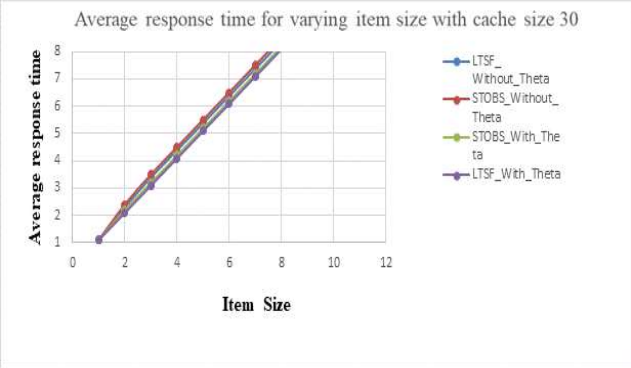
2) Performance comparison of average response time for varying cache size with different item size: Tables IV, V and Fig. 5(a), 5(b) shows the observations for average response time by varying cache size for different item size. For both STOBS and LTSF algorithms, there is significant increase of performance in network coding environments. For STOBS, average response time decreases by .18 percent for item size 60 and .16 percent for item size 30. For LTFS, average response time decreases by .17 percent for item size 30 and .16 percent for item size 60.

TABLE II. AVERAGE RESPONSE TIME WITH CACHE SIZE 30

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	1.1	1.1	1.1	1.1
2	2.3	2.4	2.2	2.1
3	3.4	3.5	3.2	3.1
4	4.4	4.5	4.2	4.1
5	5.4	5.5	5.2	5.1
6	6.4	6.5	6.2	6.1
7	7.4	7.5	7.2	7.1
8	8.4	8.5	8.2	8.1

TABLE III. AVERAGE RESPONSE TIME WITH CACHE SIZE 60

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	1.1	1.1	1.1	1.1
2	2.3	2.5	2.4	2.1
3	3.4	3.6	3.5	3.1
4	4.4	4.6	4.5	4.1
5	5.4	5.6	5.5	5.1
6	6.4	6.6	6.5	6.1
7	7.4	7.6	7.5	7.1
8	8.4	8.6	8.5	8.1



(a). Average Response Time for Varying Item Size with Cache Size 30.

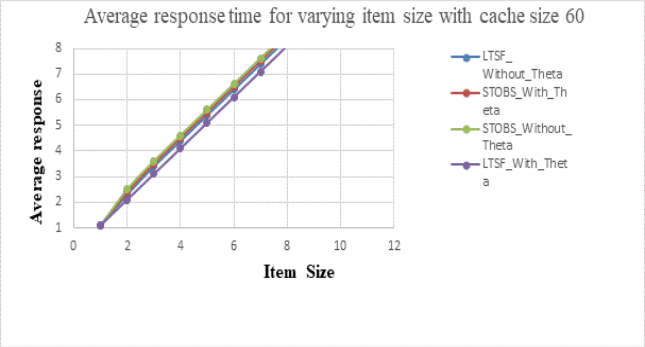


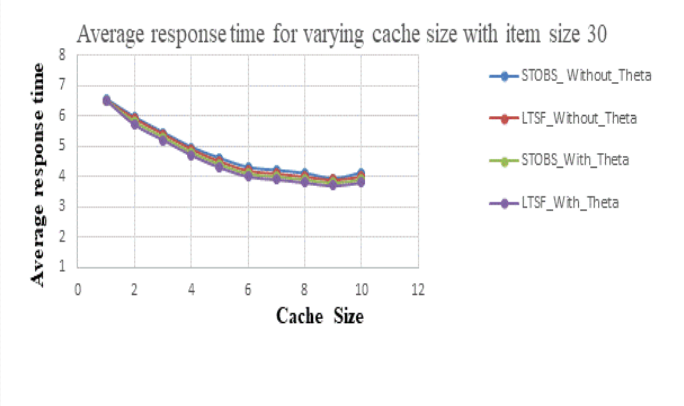
Fig. 4. (b). Average Response Time for Varying Item Size with Cache Size 60.

TABLE IV. AVERAGE RESPONSE TIME WITH ITEM SIZE 30

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	6.5	6.55	6.5	6.5
2	5.9	5.95	5.8	5.7
3	5.4	5.45	5.3	5.2
4	4.9	4.96	4.8	4.7
5	4.5	4.6	4.4	4.3
6	4.2	4.3	4.1	4
7	4.1	4.2	4	3.9
8	4	4.1	3.9	3.8

TABLE V. AVERAGE RESPONSE TIME WITH ITEM SIZE 60

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	6.5	6.5	6.5	6.5
2	5.9	6	5.8	5.7
3	5.4	5.5	5.3	5.2
4	4.9	5	4.8	4.7
5	4.5	4.6	4.4	4.3
6	4.2	4.3	4.1	4
7	4.1	4.2	4	3.9
8	4	4.1	3.9	3.8



(a). Average Response Time for Varying Cache Size with Item Size 30.

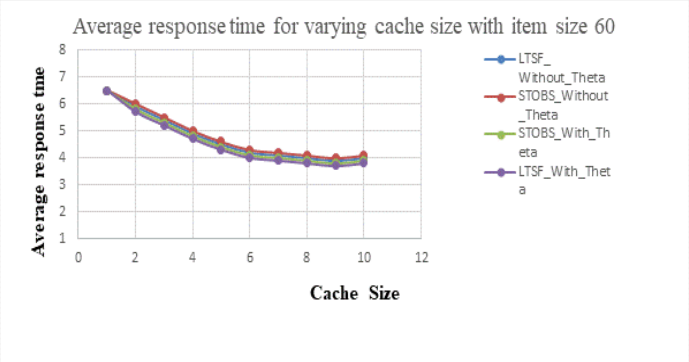


Fig. 5. (b). Average Response Time for Varying Cache Size with Item Size 60.

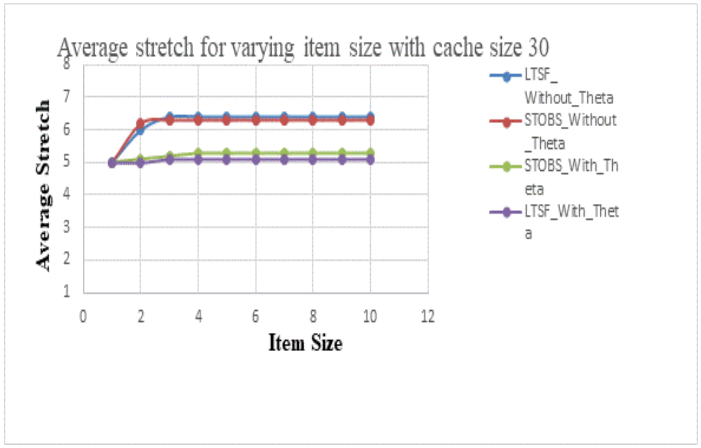
3) Performance comparison of average stretch for varying item size with different cache size: Tables VI, VII and Fig. 6(a), 6(b) shows the observations for average stretch by varying item size for different cache size. For both STOBS and LTSF algorithms, there is significant increase in performance in the network coding environment. For STOBS, average stretch decreases by .91 percent for cache size 30 and .85 percent for cache size 60. For LTFS, average stretch decreases by 1.10 percent for cache size 30 and 1.09 percent for cache size 60.

TABLE VI. AVERAGE STRETCH WITH CACHE SIZE 30

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	5	5	5	5
2	6	6.2	5.1	5
3	6.4	6.3	5.2	5.1
4	6.4	6.3	5.3	5.1
5	6.4	6.3	5.3	5.1
6	6.4	6.3	5.3	5.1
7	6.4	6.3	5.3	5.1
8	6.4	6.3	5.3	5.1

TABLE VII. AVERAGE STRETCH WITH CACHE SIZE 60

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	5	5	5	5
2	6	6.2	5.1	5
3	6.4	6.3	5.2	5.1
4	6.3	6.1	5.3	5.1
5	6.3	6.1	5.3	5.1
6	6.3	6.1	5.3	5.1
7	6.5	6.4	5.3	5.1
8	6.5	6.4	5.3	5.1



(a). Average Stretch for Varying Item Size with Cache Size 30.

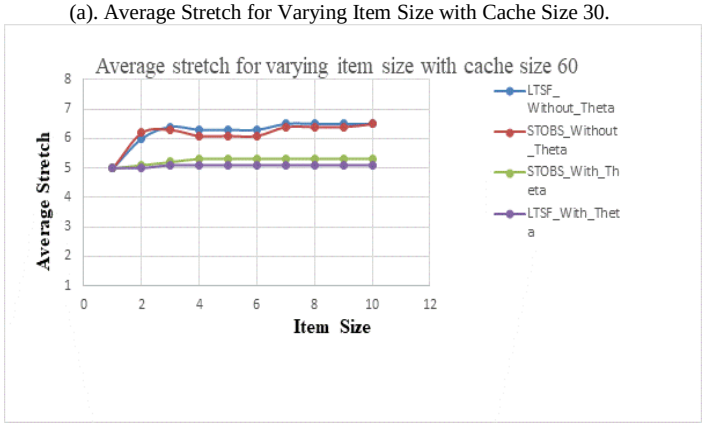


Fig. 6. (b). Average Stretch for Varying Item Size with Cache Size 60.

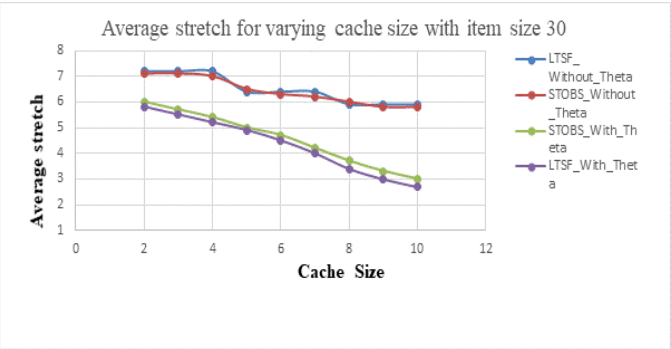
4) Performance comparison of average stretch for varying cache size with different item size: Tables VIII, IX and Fig. 7(a), 7(b) shows the observations for average stretch by varying cache size for different item size. For both STOBS and LTSF algorithms, there is significant increase in performance in the network coding environment. For STOBS, average stretch decreases by 1.75 percent for item size 30 and 2.09 percent for item size 60. For LTFS, average stretch decreases by 2.04 percent for item size 30 and 2.22 percent for item size 60.

TABLE VIII. AVERAGE STRETCH WITH ITEM SIZE 30

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	7.2	7.1	6	5.8
2	7.2	7.1	5.7	5.5
3	7.2	7	5.4	5.2
4	6.4	6.5	5	4.9
5	6.4	6.3	4.7	4.5
6	6.4	6.2	4.2	4
7	5.9	6	3.7	3.4
8	5.9	5.8	3.3	3

TABLE IX. AVERAGE STRETCH WITH ITEM SIZE 60

Serial No.	LTSF Without Theta	STOBS Without Theta	STOBS With Theta	LTSF With Theta
1	6.5	7	5	4.6
2	6.4	6.8	4.8	4.4
3	6.3	6.5	4.5	4.2
4	6.2	6.3	4.3	4
5	6.1	6.2	4.1	3.8
6	6	6.1	3.9	3.6
7	5.8	5.9	3.7	3.4
8	5.6	5.7	3.5	3.1



(a). Average Stretch for Varying Cache Size with Item Size 30.

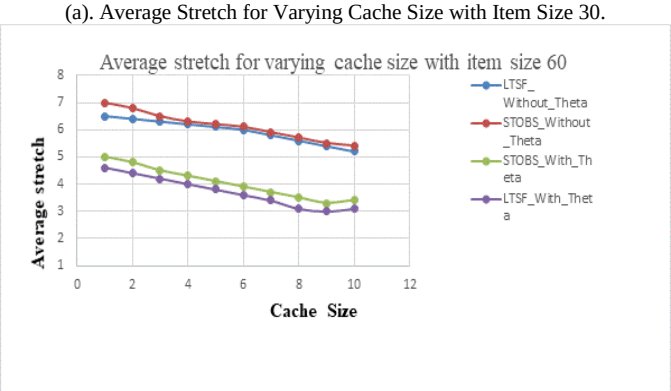


Fig. 7. (b). Average Stretch for Varying Cache Size with Item Size 60.

VI. CONCLUSION

Network coding can be widely used in the on demand wireless broadcast environment. However, it faces encoding problems while handling heterogeneous data items. It fails to provide the best possible solutions when different-sized data substance is encoded jointly. It faces a high stretch. Response time also increases when size differences of different data items are very high. Therefore, in this paper it is attempted to minimize the performance reduction difficulty of network coding in terms of heterogeneous data substance. Based on top of the generalized model proposed and discussed in, a new approach called THETA based dynamic threshold strategy has been introduced for encoding purposes. The proposed approach keeps in mind that large sized data items should not be encoded with small sized data items. The simulation results reveal interesting performance improvement of network coding. STOBS and LTSF scheduling algorithms have been used in this paper and the proposed THETA based dynamic threshold approach has been integrated with these two algorithms. With the proposed strategy, average stretch and average response time is dynamically reduced in a network coding environment. In future, other scheduling algorithms (FCFS, MRF, and LMF) can be integrated with the proposed strategy.

REFERENCES

[1] G. G. Md. Nawaz Ali, Yuxuan Meng, Victor C. S. Lee, Kai Liu and Edward Chan, "Performance Improvement in Applying Network Coding to on-demand scheduling Algorithms for Broadcasts in Wireless Networks", The Ninth International Multi-Conference on Computing in the Global Information Technology, ICCGI 2014.

[2] Yuxuan Meng, Edward Chan and Victor Lee, "Performance Simulation of Network Coding-Based on-demand Broadcast Models", IEEE, 2013.

[3] Cheng Zhan, Victor C. S. Lee, Jianping Wang, and Yinlong Xu, "Coding-Based Data Broadcast Scheduling in on-demand Broadcast", IEEE Transactions On Wireless Communications, Volume 10, No. 11, November 2011.

[4] Jun Chen, Victor C.S. Lee, Kai Liu, G.G.M.N. Ali and Edward Chan "Efficient processing of requests with network coding in on-demand data broadcast environments", ElSEVIER, 2013.

[5] Jun Chen, Kai Liu and Victor C.S.Lee "Analysis of Data Scheduling Algorithms in supporting Real-time Multi-item Requests in On-demand Broadcast Environments", IEEE, 2009.

[6] Cheng Zhan and Fuyuan Xiao "Coding based wireless broadcast scheduling in real time applications", ElSEVIER, 2016.

[7] Jun Chen,Victor C.S.Lee and Cheng Zhan "Efficient Processing of Real-time Multi-item Requests with Network Coding in On-demand Broadcast Environments", 15th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, 2009.

[8] Mohamed A. Sharaf and Panos K. Chrysanthis, "On-Demand Broadcast: New Challenges and Scheduling Algorithms".

[9] Md. Ashiqur Rahman, G. G. Md. Nawaz Ali, Yumeng Gao, Syeda K. Samantha, and Peter H.J. Chong "On Accessing Heterogeneous Data Items using Network Coding in Wireless Broadcast", IEEE, 2016.

[10] Jianliang Xu, Xueyan and Wang-Chien Lee "Time-Critical On-Demand Data Broadcast: Algorithms, Analysis, and Performance Evaluation", IEEE Transactions On Parallel and Distributed Systems, Volume. 17, No. 1, January 2006.

[11] H. Schwetman, "CSIM19: A powerful tool for building system models," in Proceedings of the 33th IEEE Winter Simulation Conference, Arlington, VA, USA, 2001.

[12] Xiao Wu and Victor C. S. Lee "Preemptive Maximum Stretch Optimization Scheduling for Wireless On-Demand Data Broadcast", International Database Engineering and Applications Symposium (IDEAS'04), IEEE, 2004.

[13] Jun Chen, Victor C.S. Lee and Kai Liu "On the performance of real-time multi-item request scheduling in data broadcast environments", ElSEVIER, 2010.

[14] Shujuan Wang, Chunting Yan and ZhengtaoY "Efficient Coding-Based Scheduling for Multi-Item Requests in Real-Time On-Demand Data Dissemination", Hindawi Publishing Corporation, volume 2016.

[15] Sachin Katti, Hariharan Rahul, Wenjun Hu, Dina Katabi, Muriel Me´dard and Jon Crowcrof "XORs in The Air: Practical Wireless Network Coding", SIGCOMM'06, September 11–15, ACM, 2006, Pisa, Italy.

[16] Tracey Ho and Desmond S. Lun "Network Coding: An Introduction".

[17] Christina Fragouli1 and Emina Soljani "Network Coding Fundamentals", Foundations and Trends in Networking, Volume 2, No. 1, 2007.

[18] YiqiongWu, JingZhao,MinShao and GuohongCao " Stretch-Optimal Scheduling for On-Demand Data Broadcasts", 2005 Springer Science + Business Media.

[19] Mohamed A. Sharaf and Panos K. Chrysanthis "On-Demand Data Broadcasting for Mobile Decision Making", Kluwer Academic Publishers, 2004.

[20] Jiun-Long Huang and Ming-Syan Chen "Dependent Data Broadcasting for Unordered Queries in a Multiple Channel Mobile Environment", IEEE Global Telecommunications Conference (GLOBECOM), November 2002.

[21] Miao Wang and Ilias Michalarias "Scheduling On-Demand Broadcast Items".

[22] Demet Aksoy and Mason Sin-Fai Leung "Pull vs Push: A Quantitative Comparison for Data Broadcast", IEEE Communications Society, Globecom 2004.

[23] Marek Konieczny "Network coding in wireless environment".

[24] M. Chaudhry, A. Sprintson, Efficient algorithms for index coding, in: INFOCOM Workshops, 2008.

[25] S. El Rouayheb, M. Chaudhry, A. Sprintson, On the minimum number of transmissions in single-hop wireless coding networks, in: Information Theory, Workshop (ITW'07), 2007, pp. 120–125.

[26] Salim Y. El Rouayheb, Mohammad Asad R. Chaudhry, and Alex Sprintson, "On the Minimum Number of Transmissions in Single-Hop Wireless Coding Networks", 5 July, 2017.

[27] Rudolf Ahlswede, Ning Cai, Shuo-Yen Robert Li and Raymond W. Yeung "Network Information Flow", IEEE Transactions on Wireless Communications, Volume. 10, NO. 11, November 2011.

[28] Xiaojiang Chen, Jingjing Zhao, Dan Xu, Shumin Cao, Haitao Li, Xianjia Meng and Dingyi Fang "Efficient Network Coding with Interference-Awareness and Neighbor States Updating in Wireless Network", Hindawi Wireless Communications and Mobile Computing Volume 2017, Article ID 4974165, 22 pages.

[29] Chen Han, Yuwang Yang and Xu Han "A fast network coding scheme for mobile wireless sensor network", International Journal of Distributed Sensor Networks 2017, Volume 13.

[30] Ali, G. G. Md. N., Lee, V. C. S., Meng, Y., Chong, P. H. J., & Chen, J. "Performance Analysis of On-Demand Scheduling with and without Network Coding in Wireless Broadcast," Future Internet, 11(12), 248, 2019.

[31] Jian Li, Tongtong Li, Jian Ren, & Han-Chieh Chao "Enjoy the Benefit of Network Coding: Combat Pollution Attacks in 5G Multihop Networks," Wireless Communications and Mobile Computing, 2018, pp. 1–13.

[32] K. Lei, S. Zhong, F. Zhu, K. Xu, and H. Zhang, "An ndniot content distribution model with network coding enhanced forwarding strategy for 5g," IEEE Transactions on Industrial Informatics, vol. 14, pp. 2725–2735, 2018.

Optimality Assessments of Classifiers on Single and Multi-labelled Obstetrics Outcome Classification Problems

Udoinyang G. Inyang[1], Samuel A. Robinson[2], Funebi F. Ijebu[3], Ifiok J. Udo[4], Chuwkudi O. Nwokoro[5]
Department of Computer Science, Faculty of Science, University of Uyo, Nigeria[1, 2, 3, 4, 5]
TEFTFUND Centre of Excellence in Computational Intelligence Research, University of Uyo, Nigeria[1, 2, 3, 4, 5]

***Abstract*—It is indisputable that clinicians cannot exactly state the outcome of pregnancies through conventional knowledge and methods even as the surge in human knowledge continues. Hence, several computational techniques have been adapted for precise pregnancy outcome (PO) prediction. Obstetric datasets for PO determination exist as single label learning (SLL), multi-label learning (MLL) and multi-target (MTP) problems. There is however no single classifier recommended to optimally satisfy the needs of all the classification types. This work therefore identifies six widely used PO classifiers and investigates their performances in all three classification categories; to find the best performing classifier. Obstetric dataset exposed to input rank analysis via Principal component Analysis, produced thirteen (13) significant features for the experiment. Accuracy, F1-measure and build/test time were used as evaluation metrics. Decision tree (DT) had an average accuracy and F1 score of 89.23% and 88.23% respectively, with 1.0 average rank. Under MLL configuration, average accuracy (91.71%) and F1 score (94.28%) were highest in the random forest (RF) which had a 1.0 average test time rank. Using MTP, DT had an average accuracy of 88.80% and average F1 score of 71.13%, the multi-layered perceptron (MLP) had the best time cost with an average rank value of 2.0. From the results, RF is most optimal in terms of accuracy and average rank value, while DT is the most efficient in terms of time cost. The comparative analysis of global averages of the six base classifiers shows that RF is the most optimal algorithm with an average accuracy of 87.3% given all three data setups in the study. MLP on the other hand had an unexpectedly high time cost, making it unsuitable for similar data classifications if time is the main criterion. It is recommended that the choice of the classifier should either be RF or DT depending on the application domain and whether or not time cost is a major consideration.**

Keywords*—Pregnancy outcome; random forest; multi-label learning; comparative analytics; machine learning algorithms; single label learning; maternal outcome prediction; decision tree

I. Introduction

Machine Learning (ML), a fast-rising branch of artificial intelligence (AI), encompasses computer science, engineering, mathematical sciences, cognitive science and many more disciplines [1]. The advancement and wide applications of ML is largely due to the availability of enormous data repositories and the satisfaction and reliability of its performances — accuracy and computational cost. It equips systems with cognitive capability of understanding the concepts of their environments through the building of models and functions, and the communication of their experiences with patterns. These models and patterns are built and implemented through the process called ML. There are two key classes of ML — supervised ML (SML) and unsupervised ML (UML) [1]. Both UML and SML draw inferences by learning, however UML utilizes datasets with input features only while SML depends on datasets having both input and target attributes for mapping and extraction of relationships between input and output feature spaces. Any dataset with target or desired output variable(s) is referred to a labelled dataset. Unlabeled datasets lack response variables therefore do not support model training activity needed by SML techniques [2-4]. In labelled datasets, every record has predefined class label(s) and supports two broad types of data mining applications — regression and classification [5]. In regression tasks, the target variable(s) is in continuous numeric form whereas classification requires class labels or categorical variables as the target. Classification is the most common and widely applied SML approach. It is aimed at identifying and assigning membership class to a new record, from a set of already defined classes [4,6]. Classification approaches are sub-divided into two groups according to the number of labels; single label and multi-label. The conventional single-label classification approach deals absolutely with disjoint classes—each record belongs exclusively to a unique class, whereas in multi-label classification the labels are intertwined and each record is associated with two or more class labels [7]. In single label problems, the categories may comprise of two labels (binary class) or more than two labels (multi-class). For example in medical diagnosis, a laboratory test result might confirm the presence or otherwise of causative organisms in the tested patient's sample while the patient can concurrently suffer from more than two diseases.

In maternal healthcare (MHC), obstetricians are confronted with the tasks ensuring safety of both the mother and baby throughout pregnancy, during delivery, and within a specified period after delivery. This is achieved by providing specialized medical care services while she is expectant, during child delivery and after delivery — antenatal, neonatal and post-natal care services. They are therefore required to obtain clinical factors for the realization of the safety of mother throughout the period during pregnancy and birth, and the newborn in a bid to minimize mortality and morbidity. These involve simultaneous predictions of multiple outcome regarding mother and neonatal status using common baseline

risk factors. Maternal outcome, mother's status during and after delivery, neonatal physiological status, conditions and overall state among others are central in MHC management. Hence, multiple target prediction, multi-label and multi-class predictions are essentially mandatory tasks in the obstetric healthcare domain. However, these maternal decisions are repeatedly made based on doctors' perceptions and experience without utilizing the pieces of vital knowledge concealed in the huge data repositories [8,9]. The author in [10] state that only about 30% of pregnancy outcomes classified by gynecologists and obstetricians concerning pathological fetus or pregnancy turns out to be true. This limitation in current medical practice has led to several complications in deliveries and avoidable deaths from the over 130 million deliveries per year globally. It is therefore expected that a robust computational technique for accurate pregnancy outcome determination will be available to assist medical personnel.

Although solutions from data mining and computational models are laudable and widely accepted methods for medical predictions, none is confirmed as a universal and best-performing model for prediction of diverse maternal outcomes; individually or in a combined target setup. This paper aims at assessing the performances and suitability on obstetrics dataset, classification algorithms under varying maternal outcome target configurations, given that they comprise binary, multi-class and multi-labeled target features. The remaining sections are structured as follows: Section 2 gives a review of related works on medical diagnoses regarding maternal health care management. In Section 3, the dataset acquisition, preprocessing and description are presented while the methodology of the comparative analytics is described in Section 4. The predictive results along with the evaluations of their performances as well as discussions are described in Section 5 while conclusions and further directions are given in Section 6.

II. Related Works

A. Single Label Learning

Classification tasks are broadly categorized into single-label learning (SLL) and multi-label learning (MLL) based on the nature of association existing between target labels and input patterns [11,12]. The goal of SLL is to build a model for the prediction of a distinct class label from a set of non-overlapping labels using input samples. It deals solely with disjoint classes and comprises two types: binary (or filtering in of textual and web-data domian) [13] and multi-class classification [11]. Binary classification has two unique class labels and involves the mapping of input features to only one of the two classes based on an explicit assessment criterion. Examples include disease diagnosis (positive or not), gender discrimination (male or female), email spam detection (spam or not), quality control (pass or fail), maternal status after delivery (alive or death) among others. Some of the famous binary classification datasets are adult dataset (adult.csv) to predict if a person's earnings per annum exceed $50,000 or not, titanic dataset (whose target has passengers who survived or not), diabetes dataset (positive or negative diabetic status), Cleveland heart disease dataset, ionosphere, banknote authentication dataset (authentic or fake). Logistic Regression, k-Nearest Neighbors (KNN), decision trees (DT), support vector machine (SVM), Naive Bayes (NB) and neural networks (NN) are some notable binary classification algorithms. Unlike binary learning problems which have two class labels, multi-class learning is applied to problems involving three or more disjoint class labels. It relies on the assumption that 1) each observation is assigned to only a single label, and 2) each class label is independent of the other [6] For example, a fruit can be one of the following types; apple, mango, orange, pear, a student can graduate with only one class of degree. Iris, zoo, waveform, dermatology, sport, MNIST, ionosphere, glass and wine datasets are some of the examples of widely used multiclass datasets that are available in data repositories and widely reported in the literature. SVM, DT, multinomial logistic regression and multi-layered perceptron are suitable algorithms for multi-class tasks. Widely adopted methodologies for multi-class tasks include; 1) decomposing target label space, via the following methods; one-vs-all, all-vs-all, and error-correcting codes 2) arrangement of the classes in a tree-like structure (hierarchical method) 3) adapting and extending binary classifiers to perform multi-class classification tasks [11,14,15].

B. Multi-label Learning

In real-world scenarios, the same set of input features are often used to concurrently predict more than one target variable. The target feature may consist of binary labels, categorical or continuous values. For binary target features the type of classification is MLL while real-valued target variables are referred to as multi-target regression. However, when the target features are categorical, it becomes a multi-target prediction problem. The MLL problem is a special kind of multi-target learning (MTL) (multi-dimensional or multi-objective), where each label can be associated with more than one values, as opposed to binary labels which have two values depicting relevance(1) or otherwise(0). Recently, MLL has progressively attracted the attention of researchers especially in ML communities and has been extensively applied to solving many problems including image and video analysis, text, bioinformatics, web mining, rule mining, information retrieval, medical diagnosis and prediction and many more [16]. Techniques advanced for MLL classification problems include; algorithm adaptation approach (AAA), problem transformation methods (PTMs) [11,12,17] and ensemble methods [11,18]. The PTMs transform the original MLL problem into multiple SLL (binary or multi-class) or regression tasks while AAAs adapt the base learning algorithms themselves to solve MLL problems rather than transforming them. PTMs adopt the basic SLL classifiers to accomplish the classification task after the transformation stage and thereafter combine the results into an MLL solution. In consideration of the flexibility of the PTMs [12,17], this work performs MLL using classifier chain (CC), bayesian classifier chain (BCC), RAndom k-labEL sets (RAkEL) and Pruned Set (PS) methods and its MTL variant Nearest Set replacement (NSR).

CCs provide a means of combining several binary classifiers into a single multi-label model that is capable of exploiting correlations among targets. It is based on binary

relevance (BR) [12,17,19] approach and beats the weaknesses of BR with an improved performance in addition to the inherited strengths of BR especially low time complexity. The main idea of CC is to incorporate label dependency to BR [7,20]. The BCC [21] uses many classifiers, one per class, linked in a chain to find a joint distribution of the classes C = (C_1, C_2, . . . , C_d) given the attributes X = (x_1, x_2, . . . , x_n). In BCC settings, a CC can be constructed by firstly inducing the classifiers that do not depend on any other class and then proceed with their descendants, according to the dependence structure which can be represented as a Bayesian network. It is an alternative method for MLL that integrates class dependencies while preserving the computational proficiency of the BR technique [21]. The RAkEL algorithm repetitively constructs a cooperative group of Label Powerset (LP) classifiers. That is, it transforms a multi-label problem into one multi-class classification problem where the possible values for the transformed class attribute is a set of distinct subsets of labels present in the original training data. Each LP classifier is trained by relying on label correlations required for ranking of the labels by averaging the zero-one predictions of each model per considered label. RAkEL offers the following advantages [13]: 1) computationally less expensive due to resulting subsets of SLL tasks; 2) improvements in the class-imbalance ratio of the dataset thereby enhancing the accuracy of minority labels; 3) collation of multiple predictions for the same label by the different LP models. The PS method leverages the most significant label relationship within a multi-label dataset by eliminating insignificant and noisy label sets which might distort the performance of the classification. This reduces the complexity originating from the label dependencies without significant information loss [20,22]. The author in [20] report from experimental evidence that the PS approach outperforms LP and other baseline methods and is highly recommended for data sets with diverse concept drifts. The NSR method is the MTL version of PS where the closest sets replace outliers, rather than using subsets.

Researchers have built and used a variety of multi-labeled datasets in disparate formats and have made them available in notable multi-label data repositories including MULAN [13], Multi-label/Multi-target Extension to Weka (MEKA), Library for SVM (LibSVM) [23], Knowledge Extraction based on Evolutionary Learning (KEEL) (Alcala-Fdez *et al*, 2011) and R Ultimate Multilabel dataset repository (RUMDR), each one using two base file formats; comma-separated values (.CSV) and attribute-relation file format (.ARFF) file formats. MULAN, scikit-multi learn, MEKA and the Multi-labelled dataset in R (mldr) package provides exploratory analysis of MLL datasets. While MEKA is a general-purpose MLL software, mldr package is limited to exploratory analysis only [24]. This work therefore adopts MEKA for MLL for comparative analytics of obstetric outcome. The degree to which samples in the dataset have more than one label of datasets (multi-labelness) is estimated with two basic parameters – label cardinality (LC) (1) and Label density (LD) (2) [24]. LC indicates the mean number of labels of the records in the dataset while LD is equivalent to LC divided by the number of labels [14,24].

$$LC = \frac{1}{n}\sum_{i=1}^{n}|Y_i| \quad (1)$$

$$LD = \frac{1}{k}\frac{1}{n}\sum_{i=1}^{n}|Y_i| \quad (2)$$

Where n represents the number of samples in the dataset, Y_i the label set of the *ith* instance, and k the sum of labels in the dataset. The LC level is directly proportional to the number of active labels per sample. Several classifiers have been developed and adapted for binary, multi-class and multi-label classification problems, but there are no classifiers recommended to optimally satisfy the needs of other classification problems. This work investigates the performances of widely used classifiers on all three types of classification with a view of finding the best performing (most suitable) one.

C. Classification Approaches for Medical Diagnostic Problems

Classification is a fundamental and pivotal task of ML and data mining (DM) applications. It is encountered in various areas, such as medicine to identify a disease of a patient, prediction of the effectiveness of surgical procedures, medical tests, and the discovery of relationships among clinical and diagnosis data. The classification of health care data (HCD) for non-faulty diagnosis and appropriate prescriptions is a rising application area of DM that is grabbing the attention of researchers [25, 26]. Several works have utilized various classification methods for diseases' diagnosis and prediction. The proper utilization of classification algorithms significantly improves the analysis, disease prediction and severity level determination in addition to ensuring early detection and effective prevention mechanisms. Over the years, analysis of morbidity and mortality data in maternal-related care has evolved from the traditional to intelligent research approaches with the aim of improving the efficiency of mother and child care during pregnancy. Nonetheless, effective analytical approaches that breed intelligent decisions are dependent on the availability of reliable data collected from the healthcare domain for the purpose of extracting knowledge for informed decision-making. This process is supported by classifiers implemented in binary, multi-class or multi-label approaches. However, a universal and multi-label classification with Extreme Learning Machine (ELM) classification approach capable of performing the functions of the three aforementioned classifiers were proposed by [11] and [14], respectively. The survey conducted by [27], provided information about the association rule, classification and cluster analysis as useful tools in the identification and discovery of risk in maternal care. These tools are developed using a few underlying algorithms that have been used for mining maternal-related care, such as DT, NB, KNN, ANN, SVM, RF, Gaussian NB and so on [28-30]. ML algorithms comprising Logistic Regression (LR), SVM, DT, BPNN, XGBoost and RF, in building predictive models for early pregnancy loss after In vitro fertilization-embryo (IVF) transfer with fetal heart rate. Each of the models experimented on the features associated with on-going pregnancy and early pregnancy loss samples. RF stood out with a high performance of 97% for recall ratio, F_1 and area under the curve (AUC), in addition to an accuracy of 99% especially for those within 10

weeks after embryo transfer. In [31] MLL was performed by adapting and extending three SLL algorithms. The comparative analysis was conducted on Genbase, Yeast and Scene datasets which were evaluated in terms of LD and LC. Genbase dataset which had 27 labels, depicts greatest multi-labelness with LD of 0.05 and LC of 1.35. Four base ML algorithms (SMO, KNN, C4.5 and NB) were used to develop a predictive model which revealed SMO as the best algorithm. However, inclusion of more well-known datasets would have helped in the comparative analysis.

The author in [28] adopted the Gaussian NB classifier-based methodology with four variables obtained from INEGI. These variables were: gender, gestational age, maternal age and fetuses. The result of the classification recorded 96% accuracy in terms of precision, recall and F1-score respectively. Similarly, the NB classifier was used to compare physician-based classification for 21,000 child and adult deaths in India, South Africa and Bangladesh. This comparative study was carried out on the classifier between two different datasets without performance evaluation of any existing analytical methods. To detect gestational diabetes mellitus (GDM) in pregnant women without a visit to the hospital, a decision support system was developed based on MLP with newly designed input [50]. The identification of predictors of in-hospital maternal mortality among women attending referral hospitals in Mali and Senegal was addressed by [51]. Nonetheless, BR, LP and CC methods with different base classifiers were used for classification [12]. Although the work was limited to the phonemes of the Tamil language only, the procedure for evaluation is useful in the classification of maternal care problems. The author in [32] compared SVM and Logistic Regression (LR) to determine their performance efficiency in pregnancy outcome prediction on anonymized dataset of 420 different pregnancy details. Four output categories were defined, and the results show that the average specificity of SVM in all four categories is at least 1% higher than that for LR, except in the case of underweight infant prediction where LR had a higher specificity. On the other hand, the average sensitivity of LR was at least 10% higher than that of SVM. The study failed to compute the classification accuracies of the designed models, although LR was adjudged as a better model. The author in [49], performed a study on the cardiotocography (CTG) dataset of the University of California Irvine machine learning repository. They compared ten machine learning algorithms; focusing on their predictive precision, recall and F1 scores. Submission of the work is that during training; DT learnt better while NB had the least learning accuracy. Conversely, between the MLP, RF, SVM, and NB algorithms; the RF had the best result with an accuracy of 92%. This is followed by MLP with 84% accuracy, then 83% for the SVM classifier with linear kernel and 77% for NB. Moreover, the work reported in [33] compared the classification ability of NB, RF, DT, and SVM on the CTG dataset using the Minimum Reduction Maximum Relevance technique for feature extraction. Their measurement matric comprised of Accuracy, Precision, Recall and F1 Score. After experiments, they report that SVM had the best classification ratings followed by RF with 96%, 88.3%, 91%, and 89.3% respectively. In addition, the work did not consider the MLP classifier even though it has been widely used with interesting results in the literature for pregnancy outcome (PO) prediction. The work reported in [10] proposed an ensemble of One Dimensional Convolutional Neural Network (1DCNN) and MLP for abnormal birth outcome detection. The study performed traced segmentation on CTU-UHB intrapartum cardiotocography dataset with 552 trace observations for class distribution equalization and 1DCNN for learning and automatic feature extraction from segmented CTG data. Classification results from the proposed model were compared with SVM, RF and MLP models trained with random weight initialization. The model evaluation using sensitivity, specificity and AUC showed that the conventional MLP classifier out-performed SVM and RF in two measures, except that it had the lowest specificity. The RF algorithm on the other hand had a higher specificity (69%) and AUC (67%) scores. SVM had 68%, 56% and 62% in sensitivity, specificity and AUC respectively, at a batch size of 500. Considering the sensitivity (80%), specificity (79%) and AUC (86%), the authors concluded that models evaluated in the study failed to produce better classification results compared to the proposed ensemble 1DCNN.

III. Data Acquisition and Feature Selection

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) features excluding the target variable. A sub-set of the attributes are; maternal age, number of children delivered, previous medical history, abortion, miscarriage, prematurity, previous illness, number of attendances to antenatal care, modal mode of delivery, antenatal registration, and mode of delivery, amongst other features. Cleaning, aggregation and pruning of attributes with only a single domain value was performed. The outcome is a dataset with thirty-five (35) input features, which were exposed to input rank analysis [34,35] via PCA in WEKA software. The selection criterion was based on eigenvalue scores not less than unity [35] regarding PO as target variable. This produced thirteen (13) significant features with a cumulative effect of 67.13%. The distribution of the variance for each factor and rank given in Table I, shows that average maternal blood pressure topped the list with an EV of 3.86 (11.7% percentage of variance), followed by average maternal weight (EV = 2.77, proportion = 8.39%). The 13th ranked attribute, average ascorbic acid level accounted for 3.17% variation with an EV score of 1.05. Target feature description of is also represented in Table I, PO consists of four Death=0) and Neonatal weight (NW) assumes low, normal or overweight as possible values.

TABLE I. RANK AND DESCRIPTION OF SIGNIFICANT INPUT ATTRIBUTES

Rank	Features	Description	EV	Proportion (%)	Cumulative (%)
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
Rank	**Features**	**Description**	**EV**	**Proportion (%)**	**Cumulative (%)**
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	-	-	-
15	Maternal status	Records whether mother is alive of death Alive=1, Death=0	-	-	-
16	Neonatal weight	Weight of the newborn low=1, normal=2 overweight=3	-	-	-

IV. MATERIALS AND METHODS

A. Predictive Analytic Models

Widely used and most performing algorithms SML algorithms; NB, SVM, DT, KNN, RF and MLP classifiers are compared. The experiment aims to observe which algorithm is capable of classifying PO in all multiple classification learning scenarios.

- KNN is a supervised classification technique aimed at predicting the target variable $y \in \{1, \dots, c\}$ given a set of features $x \in \mathbb{R}^n$ [36]. It is a type of instance-based learning, or lazy learning approach in which the approximation of functions is performed locally. KNN is based on the principle of determining a fixed number of training examples closest in distance (usually Euclidean distance) to an unknown point, and predict the label from these pieces of information. Although KNN is simple, it does not require categories to be linearly separable in addition flexibility, it is computationally costly although very fast in the training phase and arduous to estimate the optimal value of k [5,15].
- NB is a classifier based on the Bayes theorem. Results from different classification and prediction studies suggests its strength and dynamism. The implementation of NB algorithm computes the posterior probability of a hypothesis given an observed data. Given an observation c_j; NB helps determine the possibility of having d as a component of c_j, using (3):

$$P\left(C_j \middle| d\right) = \frac{P(d|C_j)\,P(C_j)}{P(d)} \quad (3)$$

where $P(d/C_j)$ is the likelihood of finding d in c_j, $P(c_j)$ is the probability of the observation c_j, while $P(d)$ is the probability of observing the data, irrespective of the specified hypothesis. The NB algorithm can often outperform more sophisticated classification methods and ranks among the topmost successful algorithms for text documents classification. It implicitly assumes that all the attributes are mutually independent which violates real-world scenarios and performs poorly on data comprising highly correlated features. It exhibits greater accuracy and speed when applied to large databases, generalizes well even with limited training samples.

- SVM is a non-parametric supervised learning classifier that finds the trade-off between minimizing the training set error and maximizing the margin for optimal classification. It is known to have the best generalization ability and resistant to overfitting [37]. It is a machine learning approach efficient for solving classification and regression problems. It relies on supervised learning models which are trained by learning algorithms and is very effective when confronted with large amount of training samples to identify patterns from them. It is one of the most powerful ML algorithms for optimization, prediction and classification tasks [38,39]. Its efficiency in the prediction of weather, power output, stock market dynamics, bioinformatics, voice and handwriting recognition, image and video analysis, and medical diagnosis, among others has been demonstrated in the literature.

The major strengths of the SVM include: 1) relatively easy training and moderate scaling even with high dimensional data; 2) trade-off between the model complexity and the error are controlled easily; 3) it can handle both continuous and categorical data as well as ability to capture the nonlinear relationships in the data; 4) assumptions regarding data structure are not required because it is a non-parametric technique; 5) provides a good generalization performance with high accuracy. Some of its weaknesses include: 1) comprehensible of results to largely depends on interpretability of the input features; 2) they are computationally costly and need a good kernel function; 3) it lacks transparency in its results because it is a non-parametric method.

- **DT** is a method for approximating discrete-valued functions, in which the learned function is represented by a decision tree. Mathematically, the i^{th} C4.5 DT classifiers solve the following problem that yields the i^{th} decision function as presented in (4).

$$f_i(x) = w_i^T \emptyset(x) + b_i \tag{4}$$

$$\text{Minimize: } L(w, \xi_j^i) = \frac{1}{2}\left||w_i|\right|^2 + C\sum_{l=1}^{N}\xi_j^i$$

$$\text{Subject to: } \underset{\sim}{y_j}\,(w_i^T \emptyset(x_j) + b_i) \geq 1 - \xi_j^i, \xi_j^i \geq 0$$

$$where\ \tilde{y}_j = 1\ if\ y_j = i\ and\ \tilde{y}_j = -1\ otherwise$$

- DT adopts hierarchical design to implement the divide-and-conquer approach. It is a non-parametric technique used for both classification and regression without functional form specification. It can be directly converted to a set of simple if-then rules to enhance human comprehensibility thereby minimizing the ambiguity of complicated decisions. DTs are effective outliers and missing values detection [5]. Because of overfitting the data, additional pruning tasks (pre-pruning and post-pruning) are required, in addition to being computationally expensive. Its performance largely depends on the characteristics of the dataset.

- RF consists of a combination of classifiers where each classifier contributes with a single vote for the assignation of the most frequent class to the input vector (x) [40]. RF is an efficient model for averaging multiple deep DT that has been trained on different parts of the same training set when the goal is to reduce variance in the result. Trees constructed with fixed training data are prone to be overly adapted to the training data. The averaging function of the RF algorithm is described in (5).

$$Y = \frac{1}{N}\sum_{n=1}^{N} Y_n\,(k') \tag{5}$$

where N is the total number of trees created in random subspaces, Y_n is the classification tree, k' represent the instance to be classified, and n is a count of the sub trees which ranges from 1 to N.

- MLP consists of multiple layers of simple, bi-state, sigmoid processing nodes of neurons that interact using weighted connections [41]. The MLP classifier is a neural network that utilizes backpropagation in prediction based on threshold functions comprising a linear combination of weight, bias, and input data, as defined by (6). Each perceptron has an activation threshold; below which the perceptron is inactivated.

$$y = \psi\,(W.X + b) \tag{6}$$

where $\boldsymbol{W}$ denotes the cumulative vector of weights, **X** is the vector of cumulated inputs, $\boldsymbol{b}$ is the bias and $\boldsymbol{\psi}$ is the non-linear activation function.

B. Problem Formulation and Dataset Modelling

The dataset on maternal outcome is modeled in three main data-setups: 1) single label/single target 2) multi-label 3) multi-target. The single label/single target setup has two variants; single target binary class (ST-BC) where each observation is only associated with a single binary class label for modeling MS target attribute; and single-target multi-class (ST-MC) representation where each instance is associated with a single target with multiple class labels (PO and NW target attributes). A record may be associated with more than two binary class labels in the multi-label (MLL) data configuration while in multi-target (MTP), every label can assume many values — nominal attributes. The input vector space $X = \mathbb{R}^k$ consists of k input variables $\{X_1, X_2, \ldots, X_k\}$ representing pregnancy risk factors for PO prediction. The target feature space $Y = \mathbb{R}^m$ has m target variables, $\{Y_1, Y_2, \ldots, Y_m\}$ for the multi-target problem. An instance (x, y), where x= $\{x_1, x_2, \ldots, x_k\}$ is the input feature vector and $y = \{y_1, y_2, \ldots, y_k\}$ is the target vector, together are constituents of X and Y respectively. The input vector space is given in (7) while (8) defines the multi target arrangement.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_{m,2} & \cdots & x_{n,k} \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} y_{1,1} & \cdots & y_{1,m} \\ y_{2,1} & \cdots & y_{2,m} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,m} \end{bmatrix} \quad (8)$$

This paper considers three maternal outcomes (m=3) as target feature; ${Y_1}^T = \{y_{1,1}, y_{2,1}, \dots, y_{n,1}\}$ is defined with an alphabet $\mathcal{Y}_1$, $\{0,1\}^n$ and is associated with the binary-class variable Maternal Status (MS). The alphabet $\mathcal{Y}_2$, $\{1,2,3\}^n$ defines the multi-class target — neonatal weight (NW) vector, represented as ${Y_2}^T = \{y_{1,2}, y_{2,2}, \dots, y_{n,2}\}$ while the vector ${Y_3}^T = \{y_{1,3}, y_{2,3}, \dots, y_{n,3}\}$ defined with alphabet $\mathcal{Y}_3$ $\{1,2,3,4\}^n$ corresponds to PO, another multi-class target arrangement. The multi-target training vector space $D = \{(x_i, y_i)\}_{i=1}^n$ is defined in (9) while the labels, $\mathcal{Y}_i \in Y$, $where\ \mathcal{Y}_i \in \{1,2,\dots,l\}$, of target variables are given in (10) – (12). Equation 13 also depicts the multi-label structure with target vector space defined over alphabet $\mathcal{Y}_4$ $\{0,1\}^n$.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_{m,2} & \cdots & x_{n,k} \end{bmatrix} \begin{bmatrix} y_{1,1} & \cdots & y_{1,m} \\ y_{2,1} & \cdots & y_{2,m} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,m} \end{bmatrix} \quad (9)$$

$$\mathcal{Y}_1 = \begin{cases} 0, & if\ an\ instance\ of\ MS\ is\ \text{"alive"} \\ 1, & if\ an\ instance\ of\ MS\ is\ \text{"death"} \end{cases} \quad (10)$$

$$\mathcal{Y}_2 = \begin{cases} 1, & if\ an\ instance\ of\ NW\ is\ \text{"}under\ weight\text{"} \\ 2, & if\ an\ instance\ of\ NW\ is\ \text{"normal"} \\ 3, & if\ an\ instance\ of\ NW\ is\ \text{"over weight"} \end{cases} \quad (11)$$

$$\mathcal{Y}_3 = \begin{cases} 1, & if\ an\ instance\ of\ PO\ is\ \text{"}miscarriage\text{"} \\ 2, & if\ an\ instance\ of\ PO\ is\ \text{"preterm"} \\ 3, & if\ an\ instance\ of\ PO\ is\ \text{"term"} \\ 4, & if\ an\ instance\ of\ PO\ is\ \text{"stillbirth"} \end{cases} \quad (12)$$

$$\mathcal{Y}_4 = \begin{cases} 1, & if\ an\ instance\ of\ a\ class\ label \\ 0, & otherwise \end{cases} \quad (13)$$

The task is to predict variants of both single-label and multi-labelled data setups. This is followed by the assessment of the weighted accuracies and computational costs of all strategies for optimal predictive power decision making in the domain of obstetric management. Table II gives the specifications of the dataset configurations. In all classification learning types, the input feature dimension is 13×1632 while the target vector for each of the SLL setting (MS, PO and NW) are column vectors. In MLL and MTL, the dimensionality of the target vector is 9 and 3 respectively, with 9 labels each. All variants of SLL setups depict LC of unity and LD of 0.5 for MS, 0.25 and 0.33 for PO and NW respectively. However, MLL and MLT have the same LC (3.00) and LD (0.33).

C. Empirical Setup

The empirical evaluation was performed on some varying experimental setups on the obstetrics outcome dataset. The different configurations were based on SLL and multi-labeled classifications types. The single labeled data configuration comprises ST-BC (where the input features are associated with one of the two class labels of the MS target) and ST-MC (where the input features are mapped to one of the more than two class labels of the PO and NW targets, respectively). All base classifiers were implemented under WEKA [42], in the SLL scenario and MEKA based frameworks [43] with the multi-labeled setting, running under Java JDK 1.7 environment. The following base classifiers: SVM, RF, DT, MLP, KNN and NB were used separately as internal classifiers in WEKA (for the ST-BC and ST-MC configurations) and MEKA (for the MLL and MTL datasets) environments. Implementations were carried out with a train/test mode of 10-fold cross validation [9] on each configuration of the dataset and repeated 20 runs with each classifier–algorithm pair on a 64bit machine of 8GB RAM size with windows 10 operating system.

The WEKA/MEKA default parameters were adopted to implement the base classifiers in both SLL and MLL settings with a batch size of 100. MLP used a learning rate of 0.3 and momentum of 0.2 while the maximum training time was 500 seconds for each iteration. There was no distance weighting associated with KNN while Linear search was used with only a single neighbor. A confidence factor of 0.25 was set for C4.5 DT. John Platt's sequential minimal optimization (SMO) algorithm was adopted for training SVM classifier with RBF Kernel function as well as epsilon value fixed at 1.0×10^{-12}. NB classifier adopted unsupervised discretization without kernel estimator. MLL and MTL setups adopted the following PTMs — classifier chains (CC), random k-label sets (RAkEL) and Bayesian classifier chains (BCC) [14] for optimality evaluations of the six base algorithms. MEKA default parameters were also adopted for the chosen PTMs and base classifiers including a batch prediction size of 100. The BCC employed CC for creating maximum spanning trees based on marginal label dependence, and NB as base classifier [21]. The RAkEL method [31] builds ensembles of Label Powerset (LP) classifiers. The training of LP classifiers relied on label correlations produced through the averaging of zero-one predictions of each model per considered label.

TABLE II. DATASET SPECIFICATIONS

Classification type		**Target feature**	**Target Vector Dimension**	**Number of Labels**	**LC**	**LD**
Single label	ST-BC	MS	1	2	1.0	0.50
	ST-MC	PO	1	4	1.0	0.25
		NW	1	3	1.0	0.33
Multi-Label	MLL	Combined (MS, PO, NW)	9	9	3.0	0.33
	MTL	Combined (MS, PO, NW)	3	9	3.0	0.33

V. Experimental Results and Discussion

A. Single Labelled Learning Results

The results for the SLL settings are presented in Tables III and IV. They represent the mean values, standard deviation (stdev) and the rank given in brackets. A rank of 1 being the highest and indicates the highest performance indicator value while a rank of 6 is the least performance rank value.

From Table III, the computed mean accuracy results show that ST-BC has the highest mean accuracy (0.950 ±0.219) and mean F1 scores (0.964±0.006). This implies that the base classifiers used in this experiment performed better in terms of accuracy and F1 score in the ST-BC configuration. In terms of classifiers, DT, RF and SVM depicts the same mean accuracy (0.964) with a slight upward variation in the *stdev* of DT. The fourth ranked classifier is MLP while NB produced the least mean accuracy (0.896 ±0.023). F1 score produced by DT in the ST-BC (0.982±0.001) is ranked the 1st while MLP yielded the smallest F1 score.

The build and test costs (Table IV), reveal that KNN is the fastest classifier with an average cost of zero during model building in all datasets, this corroborates the findings reported in [5] while MLP consumed the longest average train time in ST-BC(MS) (1.575± 0.149) and ST-MC (1.936±0.135) target setups respectively. SVM model building performance was the worst in the ST-MC (2.403±0.315).

All classifiers showed significant improvements in the testing time, DT and MLP are the top performers with average rank of 1.00 and 1.67 respectively while RF execution time was the highest time and earned a rank of 5.33. The rank of the classifiers based on accuracy and F1 score (Fig. 1) show that DT is the best ranked classifier (rank=2) in both accuracy and F1 score while SVM has a rank of 3 in both metrics. Other classifiers have an average rank greater than 3.0 in both metrics except the accuracy of RF with an average rank of 2.33. NB is the least ranked classifier based on accuracy and second lowest based on F1 score. The average rankings based on train and test time (Fig. 2) are unequal in all the classifiers. However, NB, KNN and RF ranked higher in training than testing while DT yielded the best average ranking.

TABLE III. SLL Accuracy and F1 Score (mean ± STD Deviation) and Rank (in Brackets)

Classifier	Accuracy			Average Rank
	ST-BC (MS)	***ST-MC (NW)***	***ST-MC (PO)***	
NB	0.896 ±0.023 (6)	0.807±0.027 (6)	0.784±0.025 (6)	6.0
SVM	0.964 ±0.002 (3)	0.907±0.014 (3)	0.807±0.019 (3)	3.0
kNN	0.952 ±1.195 (5)	0.909±0.067 (2)	0.738±0.059 (5)	4.0
DT	0.964 ±0.018 (1)	0.893±0.065 (4)	0.820±0.061 (1)	2.0
RF	0.964 ±0.007 (2)	0.936±0.016 (1)	0.789±0.024 (4)	2.33
MLP	0.962 ±0.067 (4)	0.887±0.016 (5)	0.813±0.022 (2)	3.67
F1 –Score				
NB	0.944±0.013 (4)	0.809±0.023 (6)	0.770±0.028 (3)	4.33
SVM	0.982± 0.001(1)	0.890±0.020 (3)	0.765±0.028 (5)	3.0
kNN	0.975±0.006(2)	0.908±0.021 (2)	0.730±0.026 (6)	3.33
DT	0.982±0.001 (1)	0.881±0.022 (4)	0.784±0.021 (1)	2.0
RF	0.955±0.011 (3)	0.930±0.019 (1)	0.766±0.026 (4)	3.33
MLP	0.943±0.004 (5)	0.865±0.021 (5)	0.774±0.026 (2)	4.33

TABLE IV. SLL Build Time and Test Time (mean ± STD Deviation) and Rank (in Brackets)

Classifier	Build Time			Average Rank
	ST-BC(MS)	**ST-MC(NW)**	**ST-MC(PO)**	
NB	0.002±0.001 (2)	0.002±0.003 (2)	0.002±0.004 (2)	2.0
SVM	0.023±0.010 (4)	1.519±0.129 (5)	2.403±0.315 (6)	5.0
kNN	0.000±0.000 (1)	0.000±0.001 (1)	0.000±0.001 (1)	1.0
DT	0.013±0.005 (3)	0.034±0.012 (3)	0.029±0.012 (3)	3.0
RF	0.300 ±0.033 (5)	0.452±0.079 (4)	0.587±0.068 (4)	4.33
MLP	1.575± 0.149 (6)	1.936±0.135 (6)	1.950 ±0.187 (5)	5.66
Test Time				
NB	0.001±0.002 (3)	0.002±0.004 (3)	0.002±0.004 (3)	3.0
SVM	0.000±0.001 (2)	0.027±0.011 (6)	0.032±0.012 (6)	4.67
kNN	0.015±0.003 (5)	0.018±0.007 (5)	0.021±0.009 (5)	5.00
DT	0.000±0.000 (1)	0.000±0.001 (1)	0.000±0.001 (1)	1.00
RF	0.009± 0.002 (4)	0.015±0.015 (4)	0.020±0.003 (4)	5.33
MLP	0.000± 0.000 (1)	0.001±0.001 (2)	0.001±0.001 (2)	1.67

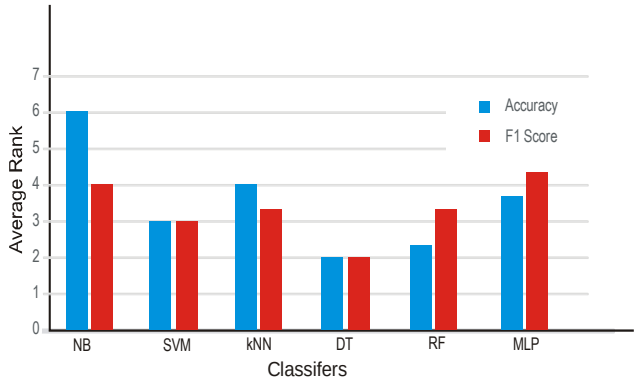


Fig. 1. Average Rank of Algorithms in SLL.

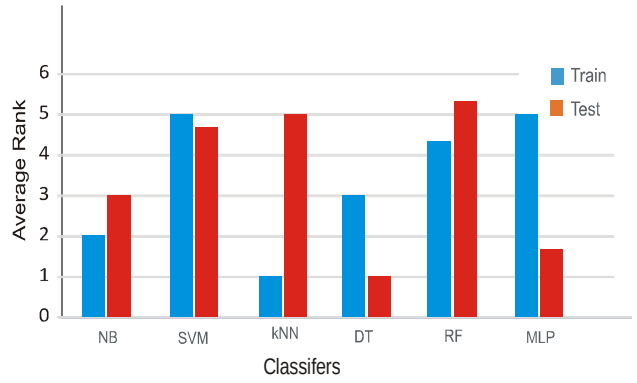


Fig. 2. Average Rank of Time Cost in SLL.

B. *Multi-labelled Learning Performance*

The distribution of average accuracy and F1 score across the PTMs and classifiers (Table V) show that NB earned the lowest accuracy and F1 score (rank=5.50) while RF produced the best performance in both Accuracy and F1 score.

It is observed that the rank of each classifier across the PTMs is the same in both metrics in addition to a marginal variation in their values. Similar results (Table VI), show that top classifies regarding accuracy earned lower ranks for time cost. Although MLP is ranked 6 with outstandingly high build time values, it competes favourably with other classifiers in the test time. KNN and DT had the best performers in the build and test times respectively while the highest execution time is exhibited by KNN followed by RF.

In the MTP scenario, Tables VII and VIII give the accuracy/F1 score and build/test time values, respectively. The F1 scores are the lowest in all dataset configurations and classification types with CC approach producing the highest average performance. The top performers are KNN, SVM and NB, in that order, and with RF having the highest average F1 score and rank of 1.25. NB earned the least rank in both accuracy (5.25) and F1 score (6.0). DT earns the highest rank (1.5) which is slightly higher than that of DT in terms of accuracy. For build and test costs, KNN utilizes an insignificant time during model build and returned as the most expensive algorithm during model execution. The reverse is the case with MLP, although the average rank of KNN is better. The ranking of DT is average in both test and build metrics while RF ranks 4.00 and 4.50 in test and build phases, respectively.

A summary of the ranks of classifiers across the datasets and classification types is given in Table IX and Fig. 3. The result shows that the ranks of classifiers in learning types varies especially between SLL and others. RF earned the best rank in MLL followed by MTP and SLL with an overall best rank of 1.78 for accuracy while depicting the worst rank in terms of time cost. DT is the second best ranked classifier regarding accuracy but is ranked the best regarding time cost while SVM is the second top classifier when considering time cost. In terms of optimality, it implies that RF is capable of producing high accuracy across dataset and classification types although is computationally expensive. This corroborates the findings reported in [9].

In term of both metrics, DT is optimal for consideration followed by KNN. It is therefore necessary to choose between RF and DT depending on the application domain and whether or not time cost should be given consideration. A cursory analysis of the result via statistical significant evaluation is presented in subsequent sections.

TABLE V. MLL ACCURACY, F1 SCORE (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	**Accuracy**				Ave Rank
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.834±0.0182 (6)	0.83 ±0.019 (6)	0.825±0.0191 (6)	0.819±0.023 (4)	5.50
SVM	0.881±0.015 (5)	0.881±0.014 (5)	0.883±0.0148 (5)	0.88±0.015 (3)	4.50
kNN	0.889±0.0162 (4)	0.89±0.016 (4)	0.885±0.0166 (4)	0.89±0.015(2)	3.50
DT	0.896±0.0152 (2)	0 .90±0.015 (2)	0.891 ±0.015 (3)	0.89±0.015 (2)	2.25
RF	0.9192±0.013 (1)	0.92±0.013 (1)	0.915±0.0132 (1)	0.914±0.014 (1)	1.00
MLP	0.894±0.016 (3)	0.894 ± 0.16 (3)	0.893±0.015 (2)	0.89±0.015 (3)	2.50
F1 Score					
NB	0.883±0.013 (6)	0.882±0.014 (6)	0.878±0.014 (6)	0.86±0.02 (4)	5.50
SVM	0.916± 0.011 (5)	0.919±0.0103 (5)	0.92±0.0104 (5)	0.92±0.01 (3)	4.50
kNN	0.922±0.012 (4)	0.923±0.012 (4)	0.920±0.0123 (4)	0.92 ±0.01 (3)	3.75
DT	0.928±0.011 (3)	0.93±0.010 (2)	0.927±0.010 (3)	0.92 ±0.01 (3)	2.75
RF	0.944±0.009 (1)	0.945±0.009 (1)	0.942 ±0.010 (1)	0.94±0.010 (1)	1.00
MLP	0.93±0.011 (2)	0.927±0.011 (3)	0.927±0.011 (2)	0.93±0.010 (2)	2.25

TABLE VI. MLL BUILD AND TEST TIMES (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	Build Time				Ave RanK
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.060±0.011 (2)	0.061±0.012 (2)	0.101±0.019 (2)	0.011±0.008 (2)	2.00
SVM	1.134±0.36 (4)	1.07±0.229 (4)	6.790 ±1.801 (4)	7.53 ±2.65 (5)	4.25
kNN	0.022±0.001 (1)	0.025±0.008 (1)	0.059±0.018 (1)	0.005±0.004 (1)	1.00
DT	0.24±0.037 (3)	0.34 ±0.062 (3)	0.847±0.13 (3)	0.140±0.028 (3)	3.00
RF	3.927±0.52 (5)	4.59±0.77 (5)	14.76±7.07 (5)	1.415±0.140 (4)	4.75
MLP	85.77±12.16 (6)	67.28 ±8.71 (6)	123.14±17.60 (6)	34.91±5.05 (6)	6.00
	Test Time				
NB	0.023±0.001 (4)	0.021±0.005 (4)	0.083±0.0157 (4)	0.029±0.006 (3)	3.75
SVM	0.008±0.015 (2)	0.005±0.003 (2)	0.025±0.024 (3)	0.047±0.05 (5)	3.00
kNN	0.913±0.13 (6)	0.756±0.120 (6)	0.90±0.2059 (6)	0.077±0.01 (6)	6.00
DT	0.002±0.002 (1)	0.002 ±0.00 (1)	0.006±0.0021 (1)	0.003±0.002 (1)	1.00
RF	0.172±0.032 (5)	0.213±0.035 (5)	0.70 ±0.37 (5)	0.039±0.010 (4)	4.75
MLP	0.012±0.004 (3)	0.009±0.004 (3)	0.0167±0.010 (2)	0.004±0.002 (2)	2.50

TABLE VII. MTP ACCURACY, F1 SCORE (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	Accuracy				Ave Rank
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.815±0.018 (6)	0.816±0.017 (4)	0.814±0.0207 (6)	0.814±0.020 (5)	5.25
SVM	0.880±0.013 (3)	0.88±0.014 (2)	0.88 ±0.014 (3)	0.88 ±0.013 (2)	2.75
kNN	0.867±0.014 (4)	0.87±0.014 (3)	0.863±0.013 (5)	0.86 ±0.014 (4)	4.00
DT	0.892 ±0.013 (2)	0.89±0.013 (1)	0.89±0.013 (1)	0.88 ± 0.013 (2)	1.50
RF	0.894 ±0.012 (1)	0.89±0.013 (1)	0.89±0.012 (2)	0.88 ±0.012 (3)	1.75
MLP	0.885±0.013 (3)	0.89±0.013 (1)	0.88±0.013 (4)	0.88 ±0.0135 (1)	2.25
F1 score					
NB	0.600 ±0.032 (6)	0.60±0.032 (6)	0.58±0.035 (6)	0.59±0.036 (6)	6.00
SVM	0.69±0.032 (4)	0.69±0.033 (4)	0.69±0.033 (4)	0.69±0.033 (4)	4.00
kNN	0.648±0.032 (5)	0.65±0.03 (5)	0.64±0.032 (5)	0.64 ±0.032 (5)	5.00
DT	0.715±0.034 (2)	0.72±0.04 (1)	0.702±0.031 (3)	0.70 ±0.035 (2)	1.75
RF	0.716±0.031 (1)	0.72±0.028 (2)	0.704±0.031 (1)	0.705±0.030 (1)	1.25
MLP	0.701±0.031 (3)	0.70±0.03 (3)	0.703±0.032 (2)	0.70 ±0.034 (3)	2.75

TABLE VIII. MTP BUILD AND TEST COSTS (MEAN ± STD DEVIATION) AND RANK (IN BRACKETS)

Classifier	Build Time				Ave. Rank
	CC	***BCC***	***RAkEL***	***PS/NSR***	
NB	0.024 ±0.008 (2)	0.032±0.025 (2)	0.014±0.022 (2)	0.01±0.0047 (2)	2.00
SVM	1.044±0.221 (4)	0.95±0.22 (4)	6.85 ±2.414 (5)	7.141 ± 2.59 (5)	4.50
kNN	0.009±0.004 (1)	0.010±0.004 (1)	0.006±0.003 (1)	0.004 ±0.003 (1)	1.00
DT	0.26 ± 0.050 (3)	0.242±0.046 (3)	0.149 ±0.032 (3)	0.139 ±0.031 (3)	3.00
RF	3.162 ± 0.52 (5)	3.60±0.911 (5)	2.180±0.43 (4)	2.048 ± 0.33 (4)	4.50
MLP	27.89 ±3.78 (6)	25.70±3.41 (6)	35.07 ±4.92 (6)	348.40±15.92 (6)	6.00
Test Time					
NB	0.012±0.004 (4)	0.014±0.003 (4)	0.029±0.006 (3)	0.034±0.007 (3)	3.5
SVM	0.003±0.008 (3)	0.004±0.010 (3)	0.049 ±0.017 (5)	0.059 ±0.049 (4)	3.75
kNN	0.34 ±0.062 (6)	0.28±0.046 (6)	0.036±0.034 (4)	0.078±0.015 (6)	5.5
DT	0.001±0.001 (1)	0.001±0.002 (1)	0.073±0.015 (6)	0.007±0.003 (1)	2.25
RF	0.114 ±0.026 (5)	0.131±0.039 (5)	0.002± 0.002 (1)	0.064±0.059 (5)	4.00
MLP	0.003 ±0.001 (2)	0.003±0.001 (2)	0.004±0.0013 (2)	0.0084±0.003 (2)	2.00

TABLE IX. AVERAGE RANKINGS (AR) OF ALGORITHMS OVER TWO METRICS AND CLASSIFICATION TYPES

	Accuracy/F1 score					Build and Test Time			
Classifier	***SLL AR***	***MLL AR***	***MTP AR***	***Global AR***		***SLL AR***	***MLL AR***	***MTP AR***	***Global AR***
NB	5.17	5.50	5.63	5.43		2.50	2.88	2.75	2.71
SVM	3.00	4.50	3.38	3.63		4.84	3.63	4.13	4.20
kNN	3.67	3.63	4.50	3.93		3.00	3.50	3.25	3.25
DT	2.00	2.50	1.63	2.04		2.00	2.00	2.63	2.21
RF	2.83	1.00	1.50	1.78		4.83	4.75	4.25	4.61
MLP	3.84	2.38	2.5	2.91		3.34	4.25	4.00	3.86

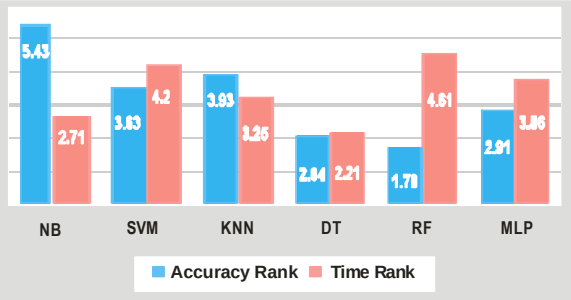


Fig. 3. Global Rank of Classifiers based on Accuracies and Time Cost.

C. Statistical Significance and Rank Validation

The main goal is to ascertain if there is any base classifiers whose performance is significantly different from others and also perform multiple comparison analysis. This was achieved by implementing non-parametric procedures [44,45] individually to each of the four categories of dataset-target setups for informed statistical inferences. Friedman test — a non-parametric variant of the repeated-measures Analysis of Variance, was used to test the null hypothesis that there is no significant difference in the performances (accuracies and time costs) of the classifiers. It compares the average rankings of the six classifiers across each of the four dataset configurations, calculating test statistic which estimates the probability of the observed rankings under the null hypothesis. Nemenyi's test and Bergmann-Hommel's post-hoc procedures implemented in R produced pairwise comparisons of all algorithms. The results are presented in the following sub-sections.

1) SLL Analysis : Friedman test on the performances of the classifiers reveals that there was no statistically significant difference in the accuracies (χ^2=10.071, df =5, p=0.0732) and time cost (χ^2=8.8571, df =5, p=0.1149) of the six classifiers at 95% confidence level (CL). This implies that the null hypothesis that there is no statistically significant difference between performances of classifiers in terms of accuracies and time cost for the SLL dataset setups is accepted. Nemenyi test (Fig. 4) compared all classifiers to each other and obtained the critical difference (CD) value of 3.2853 for both accuracies and time. As shown in Fig. 4, none of the distances separating any two classifiers in terms of their accuracy and time is greater than the CD value, this confirms that the performance of every pair of classifiers is not statistically different. In both cases, DT is the best performing classifier while RF has an average rank (AR) of 2.67 and 4.33 on accuracy and time cost respectively. Although, NB has the lowest accuracy value with an average rank of 5.17 it earned an AR of 2.67 for cost, while SVM is the most computationally expensive classifier in the SLL scenario.

2) MLL Analysis: The results of accuracies (χ^2 = 36.464, df = 5, $p = 7.67 \times 10^{-7}$) and time cost (χ^2 = 10.929, df = 5, p=0.05281) for MLL target configurations signify the existence of statistically significant difference in accuracies of classifies while the average time used by each classifier does not vary significantly at 95% CL. The CD=2.7924 (Fig. 5) is returned for both accuracy and time cost. The top three performing algorithms regarding accuracy; RF, MLP and DT, do not depict statisticaly significant difference between each other while the bottom performing classifiers kNN, SVM and NB are statistically similar. NB is lowest ranked classifier in terms of classification accuracy and is significantly different from values produced by RF, MLP and DT since their respective difference in length is greater than CD (2.7924).

Although RF is the best performing algorithm as evidence by its accuracy, it is the most time consuming algorithm with an AR of 4.75 while DT consumed the smallest amount of time in all dataset configurations, followed by NB.

3) MTL Analysis : In MTL setting, the comparison of the differences in the performance accuracy of the classifiers is statistically significant at a CL of 95% while the time costs across classifiers, statistically, does vary significantly. This is as indicated by their respective p-values and chi-squared values regarding accuracy (χ^2 = 35.125, df = 5, $p = 1.42 \times 10^{-6}$) and time (χ^2 = 5.9107, df = 5, $p = 0.315$). The CD diagram (Fig. 6), depicts the results of Nemenyi test showing the statistical comparison of all classifiers against each other by ARs based on accuracy and time. Classifiers that are not connected by a bold line of length equal to CD have significantly different ARs at 95% CL. In the case of accuracy, the values of NB are significantly different from RF, DT and MLP respectively. RF has the highest AR (1.44) followed by DT (2.12) and MLP (2.69) using accuracy while DT (2.62) and RF(4.25) stand out as the best and worst algorithms respectively when considering computation time.

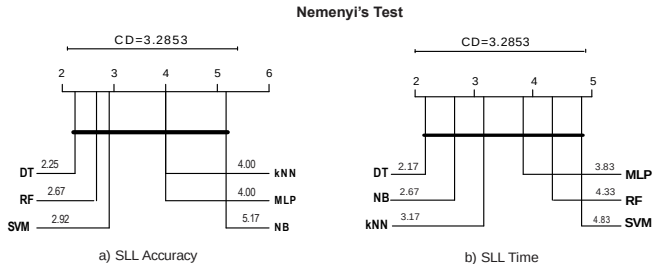


Fig. 4. CD for *Nemenyi* Test at $\alpha = 0.05$ for a) SLL Accuracy b) SLL Time.

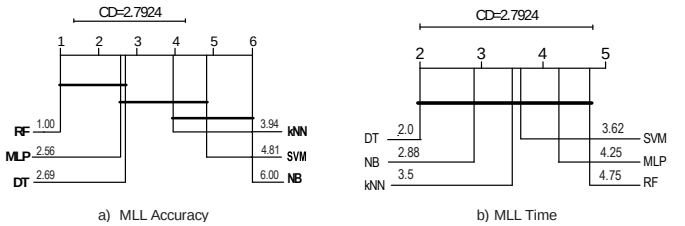


Fig. 5. CD Diagram for *Nemenyi* Test ($\alpha = 0.05$) *a) MLL Accuracy b) MLL Time.*

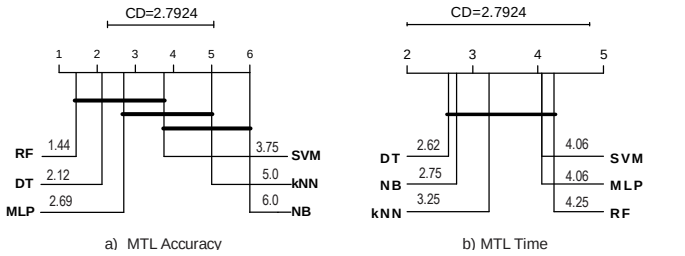


Fig. 6. CD Diagram for Nemenyi Test ($\alpha = 0.05$) a) MTL Accuracy b) MTL Time.

4) Multiple Comparison of Classifiers on all targets setups: Results of multiple comparison analysis on the combined accuracies and time costs obtained from the classifier in four dataset settings are discussed in this section. The Friedman test on aggregated values of the adopted metrics produces accuracy values (χ^2 = 70.019, df = 5, p=1.01×10^{-13}) and time values (χ^2 = 23.123, df = 5, p=3.197×10^{-4}) which depicts a statistically significant difference in performance metrics at $\alpha = 0.05$ significance level. The CD diagram (Fig. 7) obtained from the comparisons for accuracy and time, shows that the accuracy of NB significantly differs from accuracies of other classifiers while the performance of KNN differs significantly from DT and RF. The accuracy of SVM is however equivalent to others except R F and NB. RF is the highest ranked (1.61) and best performing algorithm based on accuracy followed by DT (2.36). MLP earned an AR of 3.0 and returned as the third ranking classifier while the accuracy of NB is the worst. In terms of time cost (Fig. 7b), the worst performing classifier is the RF with an AR of 4.45 and is similar to the accuracies of other classifiers except for NB and DT. DT is best classifier in terms of computational cost closely followed by NB and KNN. This implies that RF yields the highest accuracy across all classification types (dataset configuration) while it is the most computationally expensive algorithm.

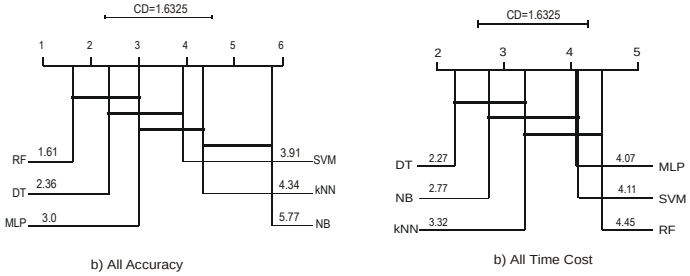


Fig. 7. CD for Nemenyi Test at $\alpha = 0.05$ for a) Accuracy b) Time.

The obtained p-values from the Freidman test specify that the null hypothesis (that all the algorithms perform the same) is reject. This, therefore, serves as the justification for conducting the post-hoc test. Bergmann–Hommel's test procedure is the most powerful, best performing, and most suitable when the number of algorithms is less than nine (9) [46-48], although it is complex and computationally expensive. Statistical pairwise comparison of the six algorithms based on average accuracies and time cost are given in Table X.

As shown in Table X, there are four major heterogeneous pairwise groupings of classifiers based on accuracy, with RF and DT being outstanding and individually significantly different from the rest of the classifiers, except MLP while NB depicts a statistically significant difference from all other classifiers. KNN-SVM and RF-DT pairs, each produced a ρ-value > 0.05, therefore statistically equivalent.

The time cost of RF is significantly different from DT (ρ<0.05) and NB (ρ<0.05) while statistically equivalent with MLP and SVM (ρ=1.0). Although the time used by DT is not statistically different from that of KNN (ρ=0.383), it exhibits a significant difference when compared with MLP (ρ=0.0110) and SVM (ρ=0.0110) in addition to RF. Pairwise comparisons involving KNN yielded no statistically significant difference as well as SVM compared with RF and MLP respectively. The summary of the Bergmann–Hommel's corrected average values (accuracy and time) of each algorithm over all the dataset is given in Table XI and Fig. 8. The results confirm that RF (accuracy=87.3%) is the best performing algorithm followed by DT (accuracy=86.3%) based on accuracy metrics while NB is the least expensive algorithm across all dataset and classification types. The ranking of classifiers considering both performance metric reveals DT (rank=2.0) as the best optimal performing classifier followed by RF (rank=3.0) while MLP (rank=4.5) depicts the worst performance.

TABLE X. CORRECTED P-VALUES USING BERGMANN–HOMMEL'S PROCEDURE FOR ACCURACY (α =0.05)

S/N	Hypothesis	ρ-value	
		Accuracy	*Time*
1	RF vs. NB	2.5×10^{-12}	2.87×10^{-2}
2	RF vs KNN	1.33×10^{-5}	3.08×10^{-1}
3	RF vs SVM	3.3×10^{-4}	1.00
4	RF vs DT	3.67×10^{-1}	1.60×10^{-3}
5	DT vs. NB	1.51×10^{-8}	1.00
6	DT vs. KNN	2.74×10^{-3}	3.83×10^{-1}
7	DT vs. SVM	2.45×10^{-2}	1.10×10^{-2}
8	NB vs. MLP	6.20×10^{-6}	1.05×10^{-1}
9	NB vs. SVM	5.72×10^{-3}	1.05×10^{-1}
10	NB vs. KNN	4.46×10^{-2}	1.00
11	MLP vs. RF	5.59×10^{-2}	1.00
12	MLP vs. KNN	6.98×10^{-2}	6.34×10^{-1}
13	MLP vs. SVM	2.14×10^{-1}	1.00
14	MLP vs. DT	5.18×10^{-1}	1.10×10^{-2}
15	KNN vs. SVM	5.19×10^{-1}	6.34×10^{-1}

TABLE XI. BERGMANN–HOMMEL'S GLOBAL AVERAGE VALUES (α =0.05)

S/N	Algorithm	Accuracy	Time	Global AR
1	NB	0.793	**0.026**	**3.5**
2	SVM	0.854	1.69	4
3	KNN	0.840	0.163	4
4	DT	**0.863**	**0.115**	**2**
5	RF	**0.873**	1.75	3
6	MLP	0.858	34.26	4.5

Fig. 8. Bergmann–Hommel's Global Rank of Classifiers based on Time and Accuracy Scores (α=0.05).

VI. CONCLUSION

Over the years, analysis of morbidity and mortality data in maternal-related care evolved from traditional to intelligent research approaches with the aim of improving the efficiency of mother and child care during pregnancy. For intelligent automated predictive solutions, ML and statistical approaches have been the most popular techniques in the literature; following the increasing clinical and administrative interest in PO determination. Results from both methods have contributed to the research of PO prediction, preconception counseling, antenatal assessment, intrapartum care, postpartum management, and reproductive health education among others. In this paper, six ML-based classifiers, including SVM, RF, DT, MLP, KNN and NB were identified as widely used and highly successful in obstetric outcome prediction. The performances and suitability of these techniques on obstetrics dataset classification under varying maternal outcome target configurations were assessed, positing that they comprise binary, multi-class and multi-labeled target features. Performance efficiency was achieved by empirical evaluation of implemented non-parametric procedures individually for SLL, MLL and MTP to enable informed statistical inferences. Using SLL, three configurations including MS, PO and NW were defined, whereas the MLL and MTP evaluations both used the CC, BCC, RAkEL, PS/NSR PMTs to evaluate performance efficiency. Dataset obtained from archives of secondary healthcare facilities in Uyo, Nigeria, was reduced feature dimension of 13 x 1632. From the results, in the SLL setup, DT had the best accuracy, F1 score and test time with an average rank of 1.0. This was followed by RF in accuracy and SVM in F1 score, while MLP had the second best time cost. NB had the worst accuracy and F1 values, while the worst test time is observed in RF. In MLL, we observed DT was least expensive in terms of time cost; whereas KNN was most expensive. RF performed better with the highest accuracy and F1 scores and was followed by DT and MLP for accuracy and F1 measures, respectively. The accuracy and F1 values obtained for NB suggests that it is the least performing classifier with the MLL setup. With an average rank of 1.50, DT had the highest accuracy in the MTP setup. This was followed by RF, while NB had the worst performance. For F1-measure evaluation, RF, DT and NB had the best, second and least performances respectively. The comparative analysis of global averages of the six base classifiers shows that RF is the most optimal algorithm with an accuracy of 87.3% given all three data setups in the study. The pole position of RF in terms of accuracy measure is in agreement with the submission in [49] (Hoodbhoy et al., 2019) that compared ten machine learning algorithms on PO determination and observed RF had an accuracy of 92% compared to lower scores obtained by MLP, SVM and NB. It also corresponds with the result obtained in [33] where the accuracy of RF was best with a score of 96%, and the work of [9]. In terms of time cost, NB is the least expensive algorithm even though it has the poorest global accuracy score. MLP on the other hand had an unexpectedly high time cost, making it unsuitable for similar data classification if time is the main criterion. Finally, from the comparative analysis, it is recommended that the choice of classifier should either be RF or DT depending on the application domain and whether or not time cost is a major consideration. As further research, the tuning of parameters of the base classifiers using evolutionary computing would be carried out in order to improve performance in terms of accuracy and computational cost.

ACKNOWLEDGMENTS

The funding for this research was provided by Tertiary Education Trust Fund (TETFUND), Nigeria through the Centre of Excellence in Computational Intelligence Research, University of Uyo. The authors would like to appreciate TETFund and the management of the University of Uyo for providing an enabling environment to carry out this research.

REFERENCES

[1] Soofi, Aized Amin, and Arshad Awan. "Classification techniques in machine learning: applications and issues." Journal of Basic and Applied Sciences 13 (2017): 459-465.

[2] Chegini, Mohammad, Jürgen Bernard, Philip Berger, Alexei Sourin, Keith Andrews, and Tobias Schreck. "Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning." Visual Informatics 3, no. 1 (2019): 9-17.

[3] Inyang, Udoinyang G., Uduak A. Umoh, Ifeoma C. Nnaemeka, and Samuel A. Robinson. "Unsupervised Characterization and Visualization of Students' Academic Performance Features." Computer and Information Science 12, no. 2 (2019): 103-116.

[4] Ekpenyong, Moses, Udoinyang Inyang, and EmemObong Udoh. "Unsupervised visualization of Under-resourced speech prosody." Speech Communication 101 (2018): 45-56.

[5] Mohamed, A. E. "Comparative study of four supervised machine learning techniques for classification." International Journal of Applied 7, no. 2 (2017).

[6] Silva-Palacios, Daniel, Cesar Ferri, and María José Ramírez-Quintana. "Improving performance of multiclass classification by inducing class hierarchies." Procedia Computer Science 108 (2017): 1692-1701.

[7] Ceylan, Zeynep, and Ebru Pekel. "Comparison of multi-label classification methods for pre-diagnosis of cervical cancer." graphical models 21 (2017): 22.

[8] Lashari, S A , Rosziati I , Norhalina Senan, and N. S. A. M. Taujuddin. "Application of Data Mining techniques for medical data classification: A review." In MATEC Web of Conferences, vol. 150, p. 06003. EDP Sciences, 2018.

[9] Inyang, U. G., Osang, F., Eyoh, I.J, Afolorunso, A. A., and Nwokoro, C.O., "Comparative Analytics of Classifiers on Resampled Datasets for Pregnancy Outcome Prediction" International Journal of Advanced Computer Science and Applications(IJACSA), 11(6), 2020. 494-504 http://dx.doi.org/10.14569/IJACSA.2020.0110662.

[10] Fergus P, Chalmer C, Montanez C.C, Reilly D, Lisboa P and Pineles (2019). Modeling Segmented Cardiotocography Time-Series Signals Using One-Dimensional Convolutional Neural Networks for the Early Detection of Abnormal Birth Outcomes. IEEE Transactions in Emerging Topics in Computational Intelligence, arXiv: 1908.02338.

[11] Er, Meng Joo, Rajasekar Venkatesan, and Ning Wang. "An online universal classifier for binary, multi-class and multi-label classification." In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 003701-003706. IEEE, 2016.

[12] Pushpa, M., and S. Karpagavalli. "Multi-label classification: Problem transformation methods in Tamil phoneme classification." Procedia computer science 115 (2017): 572-579.

[13] Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Random k-labelsets for multilabel classification." IEEE Transactions on Knowledge and Data Engineering 23, no. 7 (2010): 1079-1089.

[14] Venkatesan, R., and Er, M. J. (2014, December). Multi-label classification method based on extreme learning machines. In 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV) (pp. 619-624). IEEE.

[15] Chaitra, P. C., & Kumar, D. R. S. (2018). A review of multi-class classification algorithms. International Journal of Pure and Applied Mathematics, 118(14), 17-26.

[16] Zhang, Min-Ling, and Zhi-Hua Zhou. "A review on multi-label learning algorithms." IEEE transactions on knowledge and data engineering 26, no. 8 (2013): 1819-1837.

[17] Cherman, Everton Alvares, Maria Carolina Monard, and Jean Metz. "Multi-label problem transformation methods: a case study." CLEI Electronic Journal 14, no. 1 (2011): 4-4.

[18] Madjarov, G, Kocev, D., D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning, Pattern Recognition, vol. 45, pp. 3084-3104, 2012.

[19] Zhang, Min-Ling, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. "Binary relevance for multi-label learning: an overview." Frontiers of Computer Science 12, no. 2 (2018): 191-202.

[20] Júnior, Joel D. Costa, Elaine R. Faria, Jonathan A. Silva, João Gama, and Ricardo Cerri. "Pruned Sets for Multi-Label Stream Classification without True Labels." In 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2019.

[21] Zaragoza, J. C., Enrique Sucar, Eduardo Morales, Concha Bielza, and Pedro Larranaga. "Bayesian chain classifiers for multidimensional classification." In Twenty-second international joint conference on artificial intelligence. 2011.

[22] Read, Jesse, Bernhard Pfahringer, and Geoff Holmes. "Multi-label classification using ensembles of pruned sets." In 2008 eighth IEEE international conference on data mining, pp. 995-1000. IEEE, 2008.

[23] Chang, C.C., and Lin, C.J.: Libsvm: a library for support vector machines. ACMTrans. Intell. Syst. Technol. 2(3), 1–27 (2011).

[24] Charte, Francisco, María J. del Jesus, and Antonio J. Rivera. Multilabel classification: problem analysis, metrics and techniques. Springer, 2016.

[25] Tarle, Balasaheb, Rupali Tajanpure, and Suderson Jena. "Medical data classification using different optimization techniques: A survey." International Journal of Research in Engineering and Technology (IJRET) 5 (2016): 101-108.

[26] Umoh, Uduak A., and Udoinyang G. Inyang. "A FuzzFuzzy-Neural Intelligent Trading Model for Stock Price Prediction." International Journal of Computer Science Issues (IJCSI) 12, no. 3 (2015): 36.

[27] Mehta, R., Bhatt, N., & Ganatra, A. (2016). A survey on data mining technologies for decision support system of maternal care domain. International Journal of Computers and Applications, 138(10), 20-4.

[28] Jurado, I. C., Camarillo, D. R., & Acevedo, E. S. (2020). Problems in pregnancy, modeling fetal mortality through the Naïve Bayes classifier. International Journal of Combinatorial Optimization Problems and Informatics, 11(3), 121-129.

[29] Mathew, N. (2018, April). A Boosting Approach for Maternal Hypertensive Disorder Detection. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1474-1477). IEEE.

[30] Babu, T. A., & Kumar, P. R. (2018, January). Characterization and classification of uterine magnetomyography signals using KNN classifier. In 2018 Conference on Signal Processing and Communication Engineering Systems (SPACES) (pp. 163-166). IEEE.

[31] Tsoumakas, G, and Ioannis Vlahavas. "Random k-labelsets: An ensemble method for multilabel classification." In European conference on machine learning, pp. 406-417. Springer, Berlin, Heidelberg, 2007.

[32] Guidi G, Adembri G, Vannuccini S and Iadanza E (2014). Predictability of some Pregnancy Outcomes Based on SVM and Dichotomous Regression Techniques. IWAAL. Springer International Publishing, Switzerland. Pp.:163 – 166.

[33] Jayashree J, Harsha T, Anil K.C and Vijayashree (2020). Enhanced Optimal Feature Selection Techniques for Fetal Risk Prediction Using Machine Learning Algorithms, Int'l Journal of Engineering & Advanced Technology, 9(3), Pp.:4364–4370. doi: 10.35940/ijeat.C6502.029320.

[34] Inyang, U. G., and Akinyokun, O. C. "A hybrid knowledge discovery system for oil spillage risks pattern classification." Artificial intelligence Research 3(4), (2014): 77-86.

[35] Akinyokun, O. C., and Inyang, U. G. "Experimental study of neuro-fuzzy-genetic framework for oil spillage risk management." Artif. Intell. Research 2(4), (2013): 13-36.

[36] Losing, V., Hammer, B., & Wersing, H. (2016, December). KNN classifier with self adjusting memory for heterogeneous concept drift. In 2016 IEEE 16th international conference on data mining (ICDM) (pp. 291-300). IEEE.

[37] Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert systems with applications, 38(7), 9014-9022.

[38] Bhavsar H and Panchal H.M (2012), A Review on Support Vector Machine for Data Classification, International Journal of Advanced Research in Computer Engineering & Technology, 1(10), Pp.: 185 – 189.

[39] Meyer D, Leisch F, and Hornik K (2003). The Support Vector Machine Under Test, Neurocomputing, 55(2), Pp.: 169–186.

[40] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93-104.

[41] Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, classification.

[42] Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. "Classifier chains for multi-label classification." Machine learning 85, no. 3 (2011): 333.

[43] Read, Jesse, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. "Meka: a multi-label/multi-target extension to weka." The Journal of Machine Learning Research 17, no. 1 (2016): 667-671.

[44] Gardner, J., & Brooks, C. (2017, April). A statistical framework for predictive model evaluation in MOOCs. In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale (pp. 269-272).

[45] Janicka, Małgorzata, Mateusz Lango, and Jerzy Stefanowski. "Using information on class interrelations to improve classification of multiclass imbalanced data: A new resampling algorithm." International Journal of Applied Mathematics and Computer Science 29, no. 4 (2019): 769-781.

[46] Górecki, T., & Łuczak, M. (2017). Stacked Regression with a Generalization of the Moore-Penrose Pseudoinverse. Statistics in Transition, 18(3), 443.

[47] Calvo, B., & Santafé Rodrigo, G. (2016). scmamp: Statistical comparison of multiple algorithms in multiple problems. The R Journal, Vol. 8/1, Aug. 2016.

[48] García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information sciences, 180(10), 2044-2064.

[49] Hoodbhoy Z, Noman M, Shafique A, Nasim A, Chowdhury D and Hasan B (2019). Use of Machine Learning Algorithms for Prediction of Fetal Risk Using Cardiotocographic Data. International Journal of Applied and Basic Medical Research, Vol. 9, Pp.: 226 – 230. doi:10.4103/ijabmr.IJABMR_370_18.

[50] Muller, P. S., Sundaram, S. M., Nirmala, M., & Nagarajan, E. (2015). Application of computational technique in design of classifier for early detection of gestational diabetes mellitus. Applied Mathematical Sciences, 9(67), 3327-3336.

[51] Ndour, Cheikh, Simplice Dossou Gbété, Noelle Bru, Michal Abrahamowicz, Arnaud Fauconnier, Mamadou Traoré, Aliou Diop, Pierre Fournier, and Alexandre Dumont. "Predicting in-hospital maternal mortality in Senegal and Mali." PloS one 8, no. 5 (2013): e64157.

Pixel Value Difference based Face Recognition for Mitigation of Secret Message Detection

Alaknanda S. Patil[1], Dr. G. Sundari[2]
Department of ECE
Sathyabama Institute of Science and Technology
Chennai, Tamilnadu, India

***Abstract*—Data security is an important aspect of the modern digital world. Authentication is necessary for the prevention of data from intruders and hackers. Most of the existing system uses textual password which can provide only single-layer security. The textual passwords are simple but they may prone to spyware as well as dictionary attacks. Hence there is a need for a highly secure and multilayer security method. Steganography, the art of hiding the existence of a message by embedding it into another medium, can be exploited in an authentication system. Steganography has emerged as a technology that introduced steganalysis to detect hidden information. In this approach, the multimedia file is the input that is to be transferred over the media. On the transmitter side, the audio and video files are extracted. The secret audio file is embedded with an audio file using the LSB method while the face of the authenticated person is embedded into the video frame using the Pixel Value Differencing (PVD) method. At the receiver side, the face is extracted using the reverse PVD method and authenticated using the Convolutional Neural Network-based face recognition method. After authentication, the secret audio is extracted using the reverse LSB method. The results show that the MSE, RMSE, PSNR, and SSIM of 0.0000045303, 0.0021, 53.5877, and 0.9957, respectively.**

***Keywords*—Audio; Face recognition; Information Security; LSB; Steganography; Video**

I. Introduction

The digital world is evolving rapidly. It means that people are finding new ways of doing old tasks efficiently and creatively. Although it seems a boon, we should not forget that people won't stop using technology to their advantage. It makes the world more vulnerable as it grows. The authentication systems are the ones that require immediate attention [1].

Information hiding methods have been used in different applications for data security. This consists of copyright protection for digital media, watermarking, and steganography. Digital marketing supplies the framework to mark all data object copies with the owner's mark to insert copyright properties. Fingerprinting embeds a distinct signature for each customer purchasing the object [2].

The science of communicating secret data incorporate with communication channels is called Steganography. The communication media will be an image, audio, and video, etc. In the case of a cover image that could be color, grayscale, or even binary in which secret information is embedded, as a result, the stego object is obtained using embedding algorithms [3].

An eavesdropper may decrypt a cryptographic message but he does not even know that a steganographic message exists. Nowadays the issue of illegal copying of music files, books, and software is of critical significance. To solve such a problem steganography is being used, where any information would be encoded in digital media in such a way that it cannot be easily retrieved. There are several forms of steganography depending on the type of medium which is selected as the carrier and these include text, image, audio, video steganography, etc. The main objective of this approach is to provide multi-level security to the audio as well as video data of multimedia files using PVD and LSB algorithms.

This approach is divided into three steps: First, the secret audio sample is embedded into a multimedia file's extracted audio. The Least Significant Bit (LSB) approach is used for audio steganography. Second, the authorized user's facial image is embedded in the multimedia file frame using the Pixel Value Differencing (PVD) method. Third, a face recognition system is used to recognize the face of an authorized user. The database of authorized and unauthorized users is trained using Convolutional Neural Network (CNN) algorithm.

The proposed paper is prepared as follows; Section II offers an overview of audio-video steganography's recent development using different algorithms and their advantages and disadvantages. Section III presents the proposed methodology for the two-stage steganography approach. Section IV demonstrates the results qualitatively and quantitatively. Lastly, the conclusion is given in Section V.

II. Literature Survey

The literature survey of audio and video steganography is described in this section.

G Prasad et al. [4] proposed a method of hiding important information i.e. text, sound, or image which is embedded into an audio cover file using spatial domain techniques. The main aim of this system is to improve the data security of embedded secret audio files. This system uses the LSB technique to embed the secret audio. The performance of the system is evaluated using MSE, PSNR, and SNR. They suggest the further development can be managed to enhance the capacity and improve the robustness of an algorithm.

N. Taneja et al. [5] suggested that the security of the file transfer can be enhanced by combining encoding and digital signature. A digital signature was applied over these files and embedded with the audio file using the LSB technique. The experimental results show the increasing security and content of the hidden textual information. In the future, two encryption algorithms can be used in the file, and steganography applied to increase security.

Sattar B. Sadkhan et al. [6] presented an audio steganography method in which the LSB method is used to hide the data in the audio signal. With few changes, the bit index in stego can be changed reasonably. The method also generates a secret key, as the embedding threshold can hide and retrieve the data. The large amplitude samples produced a high bit index without decreasing the payload availability in the embedding process.

S. M. H. Alwahbani et al. [7] present the audio steganography and encryption approach in which the secret data is embedded into an audio file. Initially, data is encrypted by one-time padding, then the LSB method is applied in the spatial domain to embed the data into an audio file. The experimentation results show the efficiency of the algorithms and the quality of stego sound.

S. E. El-Khamy et al. [8] presented an efficient approach of steganography in the transform domain i.e. Discrete Wavelet Transform (DWT). Initially, the audio sample is spitting into different subbands i.e. detailed and approximate coefficient. Then select the detailed coefficient and repaced by the embedded encrypted image bit thresholded value. The encryption is performed by the RSA method. The pretraiend threshold value help to hide the cipher bits in the detailed component of the audio file. The result and analysis prove the robustness of the system in a noisy environment.

Lindawati et al. [9] presented the encryption method in which the secret message is hidden into an audio file of different formats like MP3 or .wav using the LSB method in the spatial domain. This method can hide .ppt, .docx and .xlsx files. This system is implemented on android phones. The performance of this system is evaluated using PSNR and it shows good PSNR for .wav and MP3 files.

Sattar B. Sadkhan et al. [10], proposed encryption using the AES algorithm. The secret data is encrypted and hide in the spatial domain using the LSB method to ensure confidentiality of the data. This is the simplest but robust technique. The experimental results show that the system is robust for maintaining data confidentiality. The future direction of this research is to encrypt the voice call between two phones.

Y. Bassil [11] suggested the approach to hide the essential data and information like audio samples to the public browser users. Authorized people could use a private browser to access hidden data within the web content. The experiments have provided an excellent way to hide confidential data and ensure that the LSB technique is not found. Important information can be hidden in website videos or image files in the future.

M. Than et al. [12] present the LSB-based data hiding technique. In this approach, the secret message is embedded into the .mp3 file using the LSB method. The traditional LSB method shows the weaken results besides noise. Hence this system proposed compression after combining Echo Hiding techniques.

Kaur N et al. [13] presents the procedures used for Discrete Cosine Transform (DCT) based Steganography and Discrete Wavelet Transform (DWT), and the authors provided various hiding practices or some undisclosed files in image jpg formats. Results were evaluated to know which procedure is good for hiding images.

Islam AUl et al. [14] presented hidden systematic efforts using the major bits of image pixels. The difference between bit number 5 and bit number 6 is measured, and if the outcome is unlike the secret data bit. Then the bit number 5 value is changed. The consequences of this investigation reveal that the projected method advances the signal-to-noise ratio. Several methods and techniques used in stegno scrutiny and spatial representation processes are evaluated. The likely imminent exploration drifts related to steganography safekeeping and substantiation are summarized.

Cheddad et al. [15] presented the skin tone information encryption method in YCbCr colorspace. There is different application which uses YCbCr colorspace such as object detection, video compression. This system first separates each channel of YCbCr and information is hidden in the Cr plane of YCbCr. Hence, part of the skin reviewed to hide the secret message. This system has low embedding capacity because the data is stored in only a single channel.

Kousik Dasgupta et al. [16] developed LSB-based video steganography. Eight hidden information bits are separated in 3,3,2 and embedded in the RGB pixel values of the cover frames respectively. This propagation pattern is taken as blue's chromatic effect on the human eye is stronger than red and green pixels. Video output isn't abandoned, so we could raise the payload. The proposed approach compares with current LSB-based steganography methods, witnessing an encouraging performance.

Sneha Khupse et al. [17] suggested an efficient video steganography system, using ROI instead of the whole frame in a frame. This approach uses human skin tone to mask the message. Morphological dilation and filling are used for skin area identification. Then, the video frames are translated to YCbCr color space, choosing the frame with the least square error for embedding. Hiding the hidden message is achieved inside the Cb portion of the specific video frame. This approach is constrained as only one video frame is considered for the embedding stage.

III. Face Recognition System

The comprehensive block diagram of the Face Recognition system is demonstrated in Fig. 1.

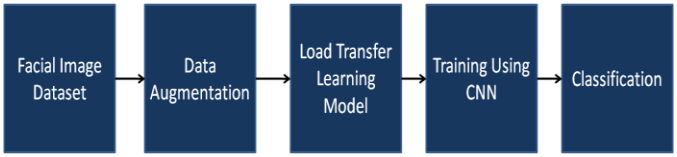


Fig. 1. Block Diagram for Embedding and Retrieval.

The key stages include Image Acquisition, Database development, Face detection, Pre-processing, Data augmentation, Feature extraction, training using CNN, and Classification. There two main phases: The training phase and the Testing (Recognition) phase.

A. Image Capture

The database is collected in real-time. The database consists of facial images of five persons in different light and luminance conditions with different angles. The collected images are in RGB format of size 227X227 pixels. The database distribution of the training and testing used for this system is tabulated in Table I.

B. Pre-Processing

Sometimes the captured facial images require a little pre-processing like cropping the face, resizing, histogram equalization for removal of illumination variance, noise reduction, thresholding, converting to the binary, or grayscale image, etc. The input image is in RGB color format. For further processing, the RGB is transformed into a grayscale image that can be attained by averaging the three-channel. Still, this method gets failure because Red color has much more wavelengths amongst these three colors, and green color has a lower wavelength than red color and soothes the eyes. Thus, we understand that there is a need to reduce the impact of red color, increase the green color's effect, and impact the blue color within these two [18]. The conversion of RGB to gray is given by.

$$Gray = 0.30 * R + 0.69 * G + 0.11 * B \quad (1)$$

R, G, and B represent the pixel intensity values of red, green, and blue pixels.

C. Data Augmentation

Artificially creating novel data from previously available training data is called data augmentation. Applying domain-specific methods to examples from training data creates new and dissimilar training examples.

Image data increase is among the most accepted data increase types. It includes creating redeveloped image forms within the training dataset that fit in the same class as the original image. The transformation consists of image manipulation operations like zooms, flips, shifts, etc.

The intention is to add new examples to the training dataset. Thus, the model is likely to observe the training set image variations.

TABLE I. DATABASE DISTRIBUTION FOR THE FACE RECOGNITION SYSTEM

Data / Labels	Total facial Images	Training Facial Images	Testing Facial Images
1	973	779	194
2	830	664	166
3	924	740	184
4	842	674	168
5	1453	1089	364

D. Training and Testing using CNN

CNN's demonstrated effective image classification. CNN consists of neurons, kernels, or filters with weights, parameters, and biases. Each filter receives inputs, executes convolution. CNN's structure contains Rectified Linear Unit (ReLU) and Fully Connected Layers (FCL).

- Convolutional Layer: The convolutional layer forms CNN's central building block, which performs the heaviest computational work. The primary aim of the convolution layer is to extract input data images. A set of learnable neurons transforms image input. It generates a function map or activation map in the output image and is then fed as input data to the next convolution sheet [19].
- Pooling Layer: The pooling layer reduces the dimensionality of each activation diagram but has the most essential details. Separate input images into non-overlapping rectangles. Each field has an average or maximum non-linear activity. This layer achieves quicker convergence, better generalization, stable translation, and modification, and is usually positioned inside convolutional layers [19].
- ReLU Layer: ReLU is a non-linear operation using the rectifier. Applied per pixel, the map reconstitutes all negative values to 0. To understand how ReLU operates, we accept an input given as x, and in the neural network image literature, the rectifier is called $f(x) = max(0, x)$. Using FCL, these features are used to identify the input image in various groups depending on the training dataset. Using the Softmax activation mechanism, FCL is called the final pooling layer inputting features to a classifier. Summing the maximum layer performance possibilities is 1. Using Softmax as an activation mechanism is verified.

E. AlexNet

AlexNet among the popular deep networks used for several computer vision applications. In this approach, the transfer learning of a trained CNN model that is AlexNet is employed for face recognition. The AlexNet model architecture is shown in Fig. 2.

AlexNet has five convolutional layers trailed by three fully connected layers. These convolutional layers extract essential features from the image. Every convolutional layer comprises linear convolution filters followed by ReLU activation, normalization, and max pooling. The primary layer is the input layer, which takes images having size 227-by-227-by-3. The very first convolution layer has 96 filters, each of which is sized 11x11x3 with pace four and no padding. The first convolutional layer results are passed on to the ReLU layer, which is followed by the max-pooling layer. The purpose behind using the ReLU activation function is the prevention of propagation of any non-positive value in the network. The pooling layer aims to lessen computation and control overfitting. The second convolutional layer comprises 256 filters sized 5x5 with pace one and padding 2. The third, fourth, and fifth convolution layer executes 3x3 convolution with rate one and padding 1. Only convolutional layers 1, 2, and 5 have

max-pooling. Three fully connected layers trail the down-sampling and convolutional layers. The final fully connected layer uses features learned from the last layer to execute the classification task. This layer is followed by a softmax layer, which will normalize the output.

In this approach, we have trained AlexNet for face recognition. Fig. 3 shows the AlexNet training. Accuracy of 99.66% is achieved during the training.

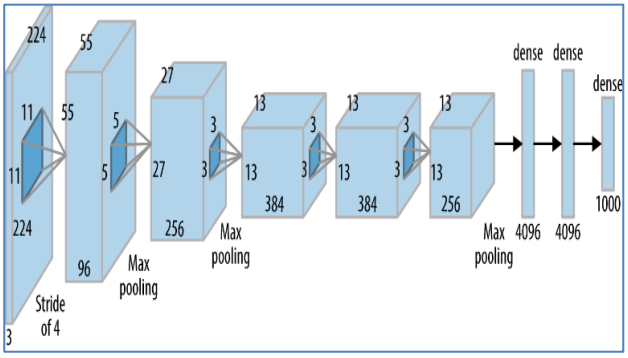


Fig. 2. Architecture of AlexNet.

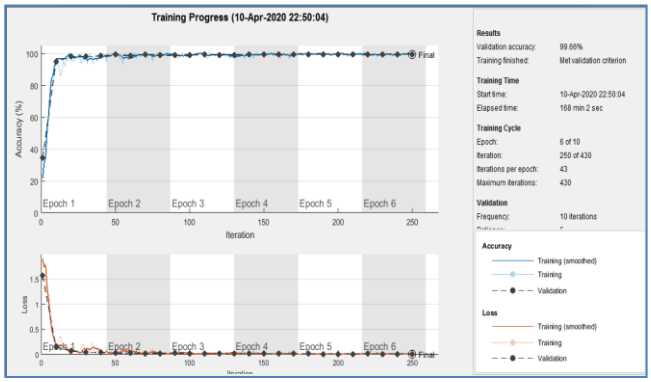


Fig. 3. Alexnet Training Progress.

IV. Proposed System

The proposed audio-video steganography system is as shown in Fig. 4. The proposed audio-video steganography system is divided into two parts transmitter and receiver. In the transmitter section, initially, the audio (called cover audio) and video (cover video) are separated from the multimedia file.

Direct Sequence Spread Spectrum (DSSS) is a spread spectrum technique whereby the secret audio data is multiplied with a pseudorandom noise (PN) spreading code. This spreading code has a higher chip rate/bit rate, which results in a wideband time-continuous scrambled audio. This processed secret audio and cover audio files embedded using the LSB method. If the bit of cover audio C(i,j) is equal to the message bit m of the secret massage to be embedded, C(i,j) will remain unchanged; if not, then set C(i,j) to m. The message embedding processing is as elaborate in Eq. 2 [20].

$$S(i,j) = C(i,j) - 1, if\ LSB\big(C(i,j)\big) = 1\ and\ m = 0$$

$$S(i,j) = C(i,j), if\ LSB\big(C(i,j)\big) = m$$

$$S(i,j) = C(i,j) + 1, if\ LSB\big(C(i,j)\big) = 0\ and\ m = 1 \quad (2)$$

where LSB($C(i,j)$) be the LSB of cover audio, $C(i,j)$and m is been the next message bit which is to be embedded, $S(i,j)$ is the stego audio. The output file is called a stego audio file. The secret audio embedding process is as follows.

1) First, extract the bit from the cover audio.
2) Second, extract the bit from the secret audio
3) Choose the first bit, pick the secret audio and place it in the first component of the bit
4) Place a terminating symbol to indicate the key end. This algorithm used 0 as a terminating symbol.
5) Insert some secret audio file in each first component of the next bit, replacing it
6) Repeat step 6 till all the bit of secret audio has been embedded.
7) Place some terminating symbols to indicate data end.
8) Output stego audio.

In another step, the authorized user's facial image is embedded with the extracted frame of video using the PVD method. The insertion process is explained below.

1) For each sequential pixel ($P_{(i,x)}$and $P_{(i,y)}$) in the cover image, calculate the difference between $P_{(i,x)}$and $P_{(i,y)}$ as d_i. find the lower limit (I_j) and the higher limit (u_j) from the range table (R_j) based on the d_i value.
2) Calculate $w_j = u_j + I_j + 1$
3) Calculate the value of $t_i = \log(w_j)$ with the log base of 2.
4) t_i value determines how many bits can be inserted.
5) Take the t_i message, i_θ is the decimal value of t_i
6) Calculate the value $\delta_i = \theta_i + I_j$
7) Calculate the value of $m = abs(\delta_i - d_i)$
8) Calculate $p'_{(i,x)}$and $p'_{(i,y)}$ by using Eq. 3.

$$p'_{(i,x)}, p'_{(i,y)} = \begin{cases} \left(P_{(i,x)} + \left[\frac{m}{2}\right], P_{(i,y)} - \left[\frac{m}{2}\right]\right), \\ P_{(i,x)} \geq P_{(i,y)} and\ d'_i > d_i; \\ \left(P_{(i,x)} - \left[\frac{m}{2}\right], P_{(i,y)} + \left[\frac{m}{2}\right]\right) \\ P_{(i,x)} < P_{(i,y)} and\ d'_i > d_i; \\ \left(P_{(i,x)} - \left[\frac{m}{2}\right], P_{(i,y)} + \left[\frac{m}{2}\right]\right) \\ P_{(i,x)} \geq P_{(i,y)} and\ d'_i \leq d_i; \\ \left(P_{(i,x)} + \left[\frac{m}{2}\right], P_{(i,y)} - \left[\frac{m}{2}\right]\right) \\ P_{(i,x)} < P_{(i,y)} and\ d'_i \leq d_i; \end{cases} \quad (3)$$

Using the above PVD method, a stego facial image is obtained. The stego audio and stego video are then embedded and transfer over the communication channel with additive white noise.

1) For each successive pixel in the stego image i.e. $p'_{(i,x)}$ and $p'_{(i,y)}$, determine the difference of $p'_{(i,x)}$and $p'_{(i,y)}$ as d'_i. Find the lower limit (I_j) and the higher limit (u_j) from the table range (R_j) based on the value of d'_i,.
2) Determine $w_j = u_j - I_j + 1$.
3) Calculate the value of $t_i = \log(w_j)$ with log_2.
4) t_i the value determines how many bits can be inserted.

5) Calculated $d_i^{"} = d_i' - l_j$, convert $d_i^{"}$ into binary values with the length of t_i

6) The conversion of $d_i^{"}$ into binary with the length of t_i is the hidden message.

7) Differential value algorithm of 2 pixels: finding a separate value of two adjacent pixels, two-pixel width difference, converting d to binary with t length, bit message length,

message power inserted in t in a bit, converting inserted messages to decimal, measuring two new pixels after inserting messages.

On the receiver side, the white noise from the stego multimedia file is removed. The decryption process is started with the separation of stego audio and stego video from the noiseless multimedia file. First, the face from the stego video file is extracted using the reverse PVD method. The process is explained.

The transmitted face is extracted from the above reverse PVD process, which is the face recognition process. The CNN algorithm is used to recognize the face of an authorized user. The extracted face is tested with a trained face recognition model, which gives us whether the user is authorized. The secret audio from stego audio proceeds if the user is authorized. The secret audio from stego audio is extracted using the reverse LSB method. The extraction process is as follows.

1) Extract the bit of the stego audio.

2) Now, start from the first bit and extract the stego bit from the first component.

3) Repeat step 2 till all the bit of secret audio has been extracted.

4) Obtained secret audio.

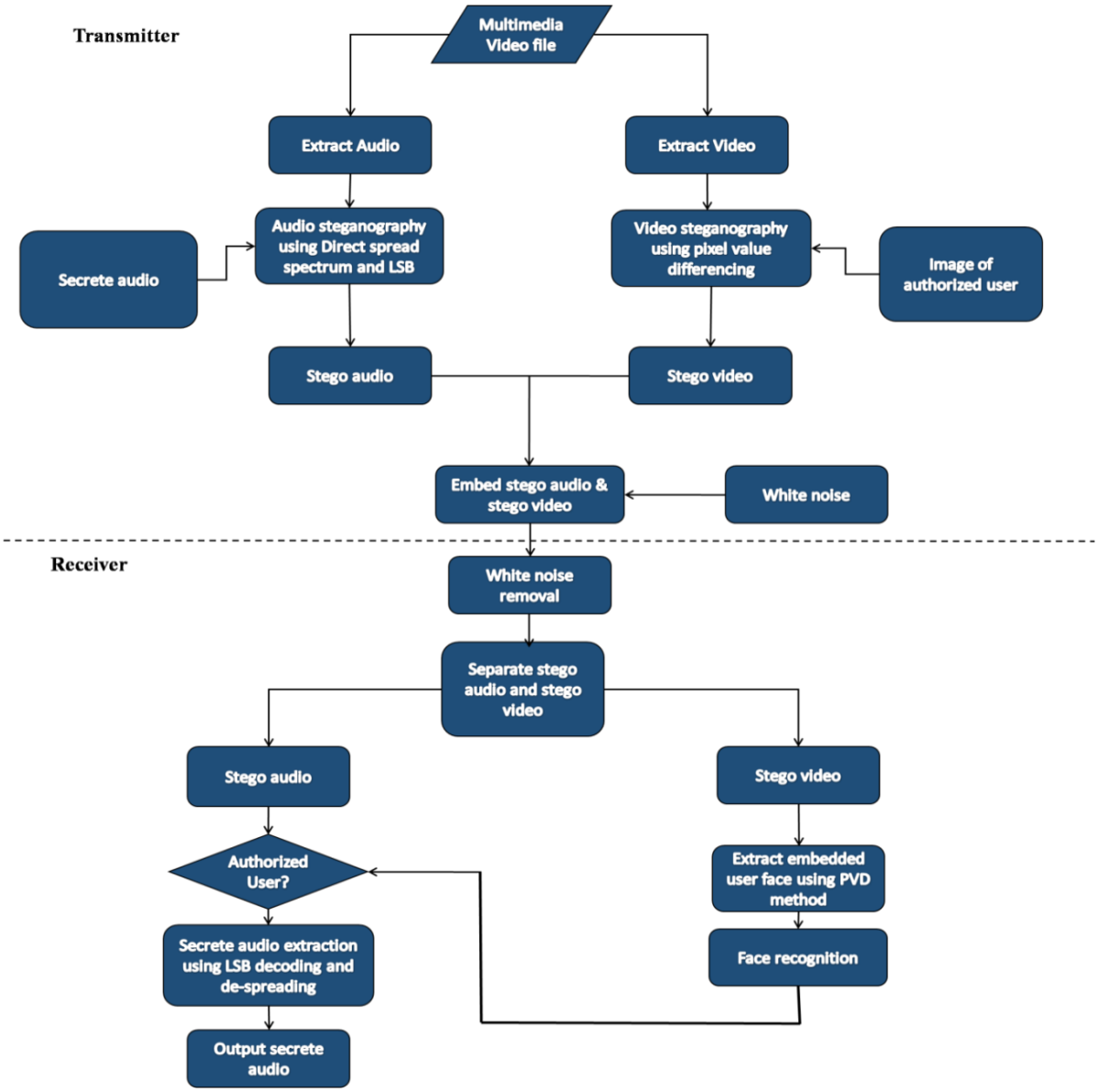


Fig. 4. Block Diagram of Proposed Audio Video Steganography.

V. RESULT

The proposed system is implemented using MATLAB 2020a, the X64 bit software.

A. Analysis of Face Recognition

In this approach, the face recognition system is implemented using a deep CNN algorithm. The training parameter used to train the face recognition system is as tabulated in Table II.

TABLE II. TRAINING PARAMETER OF CNN ALGORITHM FOR THE PROPOSED FACE RECOGNITION SYSTEM

Training Parameters	**Values**
Training algorithm	*'sgdm'*
Momentum	0.9000
Batch size	10
Initial Learning Rate	3e-4
Drop Period	10
Drop Factor	0.1
Gradient Threshold Method	'l2norm'

The face recognition system is implemented using the CNN algorithm.

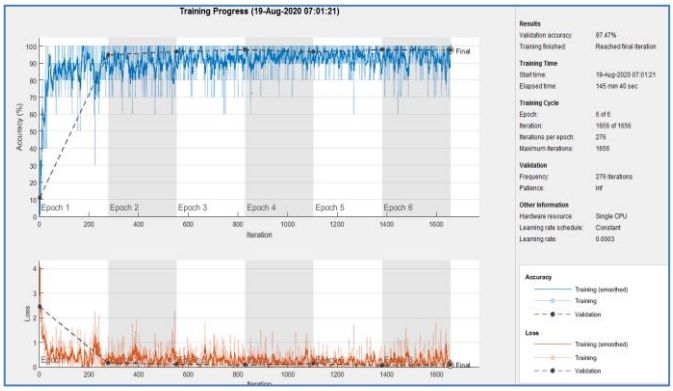


Fig. 5. Traning Progress Graph of CNN based Face Recognition System.

The training progress graph given in Fig. 5 shows that training and validation accuracy graphs follow each other and achieved a validation accuracy of 97.47%. The trained model is portable and used to recognize the authorized and unauthorized user with high accuracy. The qualitative analysis of the face recognition system is, as shown in Fig. 6.

The proposed system's qualitative analysis shows that the proposed system accurately classified the facial samples into authorized and unauthorized users. Fig. 6(a-d) are the authorized samples, while Fig. 6(e) is the unauthorized user's sample face, which is provided to the trained CNN model's input. The CNN model gives accurate output for each sample.

B. Analysis PVD based Video Steganography

In this method, the video frame is considered the cover image, and the face image of the authentic user is regarded as a secret image. The video steganography aims to hide the video frame's facial image, which is not visible to the intruder. The results of the PVD-based video steganography are as shown in Fig. 7.

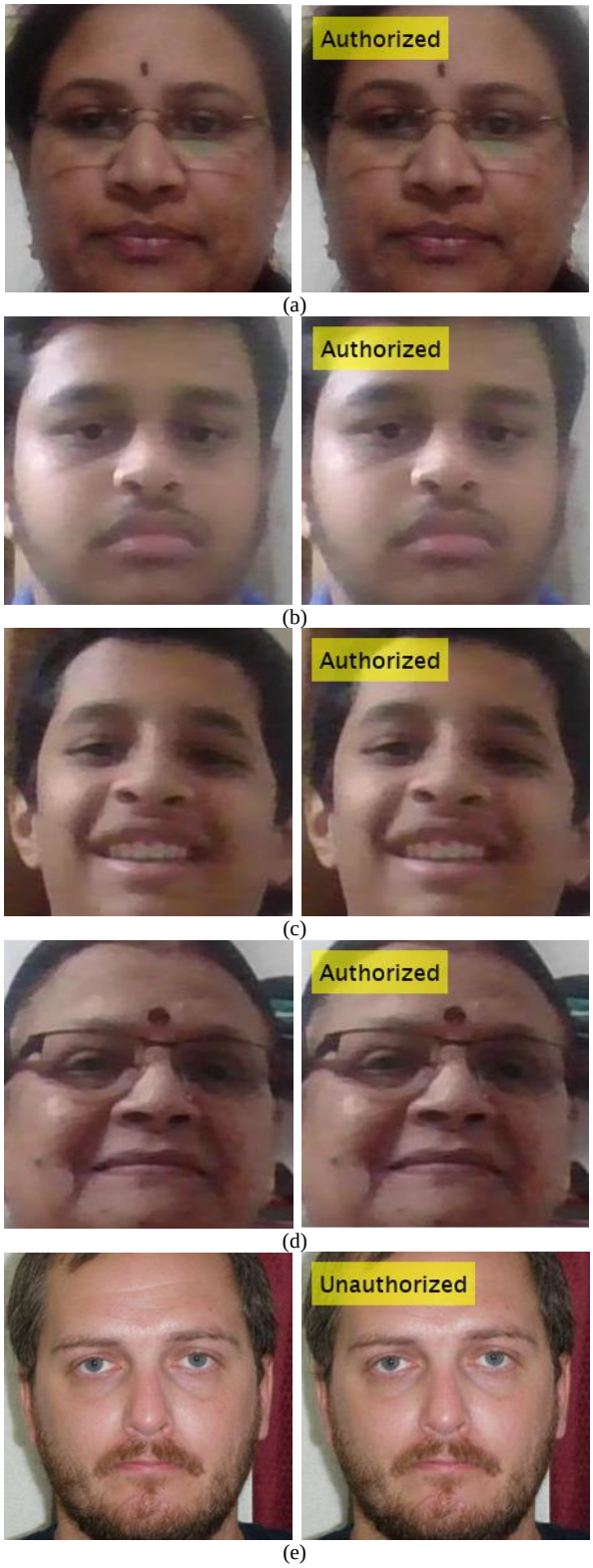
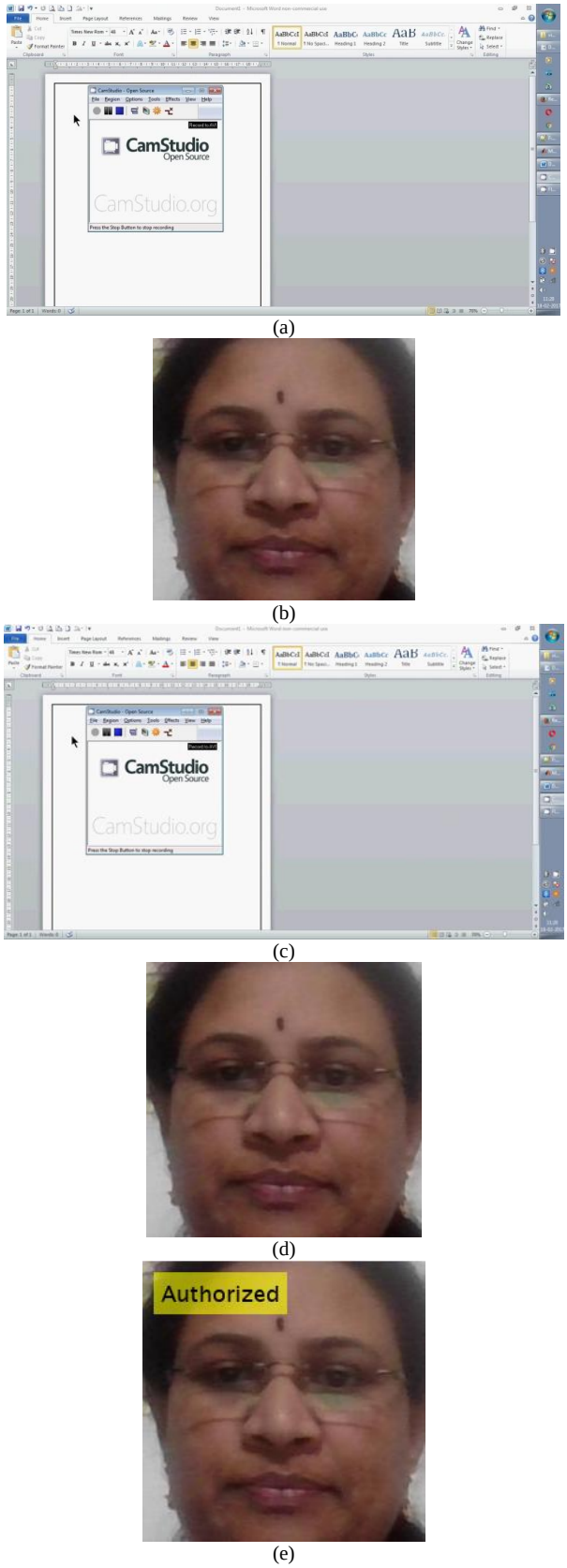


Fig. 6. Qualitative Analysis of Face Recognition System (a)-(d) Authorized User (e) Unauthorized user.



(a) (b) (c) (d) (e)

Fig. 7. Results of PVD based Video Steganography (a) Cover Frame (b) Secret Image (c) Stego Frame (d) Decrypted Secret Image (e) Authorized Face Recognition Output.

Fig. 7 shows the results of the PVD-based video steganography. The cover image and secret image are shown in Fig. 7(a) and Fig. 7(b). The authorized user's secret image is embedded in the cover image using the PVD method and created a stego image shown in Fig. 7(c). Fig. 7(c) shows that the cover image and stego image are visually similar. Hence the intruder cannot predict the embedded secret message with naked eyes. The reverse PVD method is used at the receiver side to extract the face of the authorized user, shown in Fig. 7(d). Finally, the extracted face is tested with a CNN-trained model to predict the authorized person, shown in Fig. 7(e).

C. Analysis LSB based Audio Steganography

In this section, the LSB-based audio steganography process is presented with the analysis shown in Fig. 8. The secret audio and the extracted cover audio extracted from the video multimedia file are presented in Fig. 8(a) and Fig. 8(b). The secret audio and cover audio are embedded using the LSB method called stego audio, presented in Fig. 8(c). From waveform analysis, it is observed that the cover audio waveform and stego audio waveform are looking visually similar. Therefore, it is a problematic intruder to recognize the embedded secret audio. In the decrypted process, the stego audio is extracted using the reverse LSB method shown in Fig. 8(d).

D. Quantitative Analysis

The Results of the systems are evaluated using Peak Signal to Noise Ratio (PSNR), Root Mean Square Error (RMSE), and Structural Similarity Index Matrix (SSIM). The detailed explanation of this parameter is as explained as follows.

- PSNR: PSNR is the parameter of the audio file that means Peak Signal to Noise Ratio. PSNR and MSE both are inversely proportional to each other, and the following equation can measure PSNR.

$$PSNR = 10\log_{10}\left[\frac{I^2}{MSE}\right] \quad (2)$$

Where I is the maximum possible value of audio.

- RMSE: RMSE is a parameter that means Root Means Square Error, calculated as the square root of MSE.

$$RMSE = \sqrt{\frac{1}{[N\times M]^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(X_{ij}-Y_{ij}\right)^2} \quad (3)$$

- SSIM: SSIM is the measure of the quality degradation caused by the modification and loss in the data transmission. The SSIM is calculated in this approach is between the original audio and extracted audio.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}C+_2)}{(\mu_x+\mu_y+C_1)(\sigma_x+\sigma_y+C_2)} \quad (4)$$

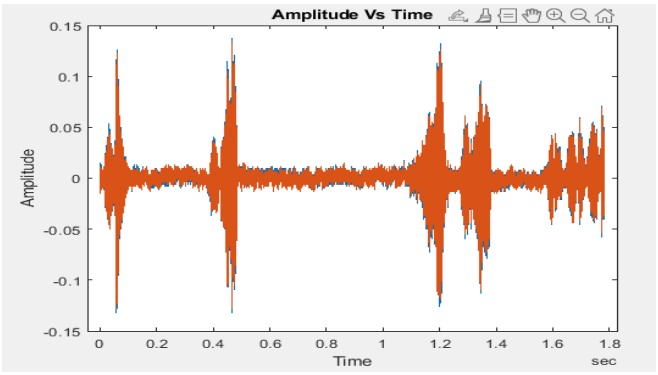
where μ_x, μ_y, are the local mean, σ_x, σ_y are the standard deviation and σ_{xy} is the cross-covariance for data x, y.

The mean, standard deviation, and cross variance is given by

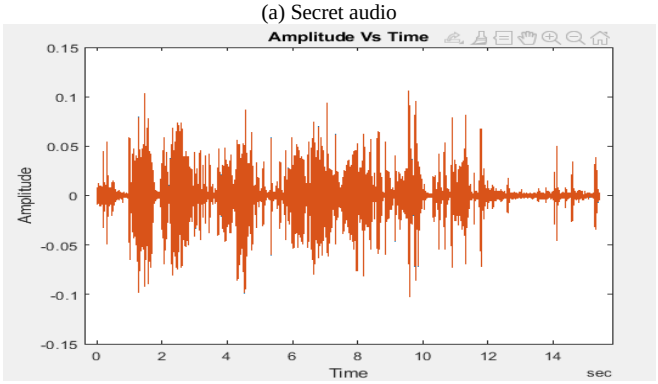
$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (5)$$

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{\frac{1}{2}} \quad (6)$$

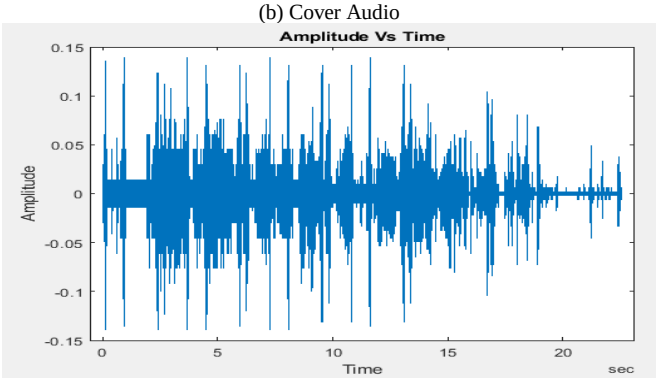
$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \quad (7)$$



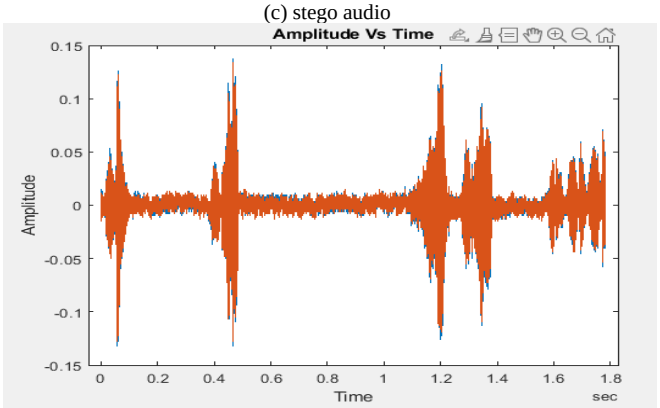
(a) Secret audio



(b) Cover Audio



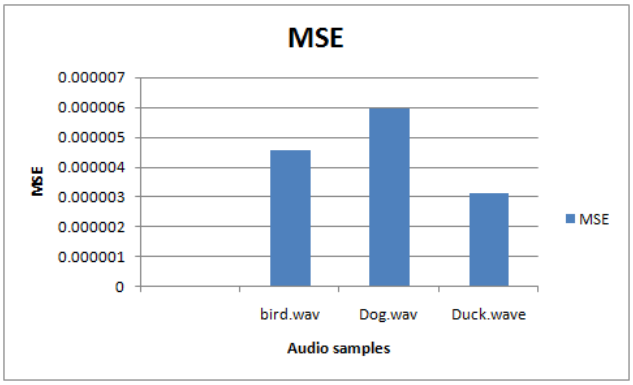
(c) stego audio



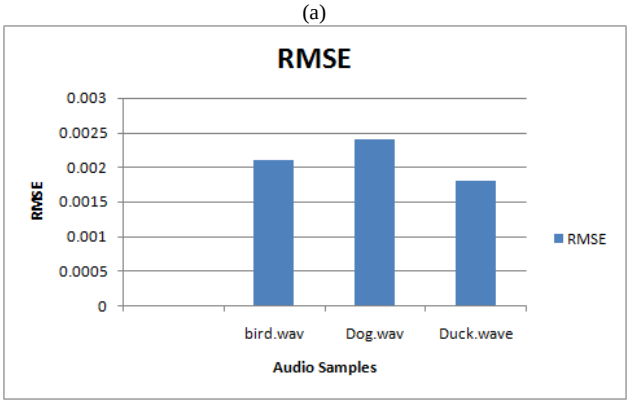
(d)

Fig. 8. Results of LSB based Audio Steganography (a) Secret Audio (b) Cover Audio (c) Stego Audio (d) Decrypted Secret Audio.

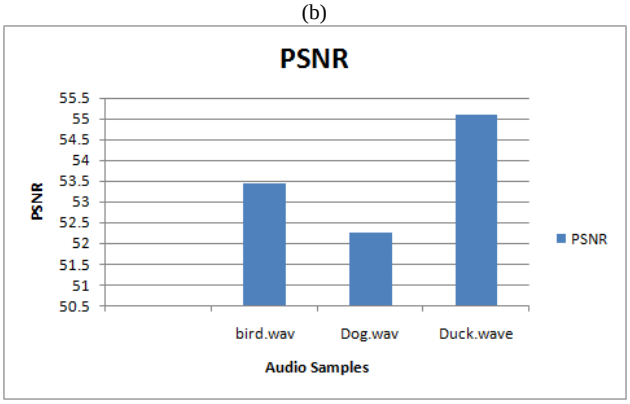
The qualitative analysis in terms of MSE, RMSE, PSNR, and SSIM of the audio steganography is shown in Fig. 9(a-d).



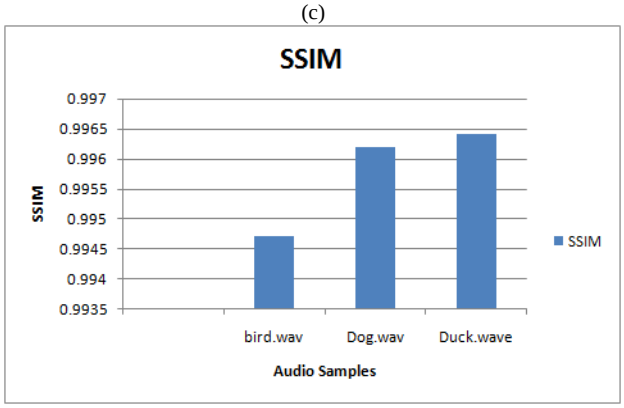
(a)



(b)



(c)



(d)

Fig. 9. Graphical Analysis of the Audio Steganography (a) MSE (b) RMSE (c) PSNR (d) SSIM.

The graphical analysis shows that the MSE and RMSE values of the original secret audio and extracted secret audio are minimal. In contrast, PSNR and SSIM values are high. Hence it can be concluded that the system is highly precise and can be used for steganographic applications.

VI. Conclusion

In this paper, the Audio and video steganography approach is presented. The video steganography is performed using a pixel value differencing method while audio steganography is performed using the least significant method. The authorized user is recognized using the CNN algorithm, which shows excellent validation accuracy of 97.47%. The Results of the system are presented using MSE, RMSE, PSNR, and SSIM. The results show that audio and video steganography leads to promising results. This method could widely be used to modify LSB's without hampering the audio quality of the sound. The proposed approach attained enhanced MSE, RMSE, PSNR, and SSIM of 0.0000045303, 0.0021, 53.5877, and 0.9957, respectively.

In the future, new data-hiding schemes will be worked to improve the embedding capacity by merging the PVD scheme and hidden sharing scheme.

References

[1] N. Kaushik, P. Sultana H, S. Jayavel, "Remote Authentication using Face Recognition with Steganography", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018, pp. 351-354.

[2] Yusuf Perwej, Firoj Parwej, Asif Perwej, "An Adaptive Watermarking Technique for the copyright of digital images and Digital Image Protection", The International Journal of Multimedia & Its Applications (IJMA), April 2012, Volume 4, Number 2, Pp. 21-38.

[3] Nagham Hamid, Abid Yahya, R. Badlishah Ahmad & Osamah M. Al-Qershi, "Image Steganography Techniques: An Overview", International Journal of Computer Science and Security (IJCSS), Volume (6): Issue (3): 2012 168-187.

[4] G. Prasad TVS and S. Varadarajan, "A Novel Hybrid Audio Steganography for Imperceptible Data Hiding," IEEE 978-1-4799-8081-9/15/$31.00 ©, 2015.

[5] N. Taneja and P. Gupta, "Implementation of Dual Security through DSA and Audio Steganography," International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, India, 2015.

[6] Sattar B. Sadkhan, Dr. Nidaa A. Abbas, "Multidisciplinary Perspectives in Cryptology and Information Security", Book, Publisher IGI Global, 2014.

[7] S. M.H. Alwahbani and H. T.I. Elshoush, "Chaos-Based Audio Steganography and Cryptography Using LSB Method and One-Time Pad," Springer International Publishing AG 2018 Y. Bi et al. (eds.), Proceedings of SAI Intelligent Systems Conference (IntelliSys), 2016.

[8] S. E. El-Khamy, N. O. Korany, and M. H. El-Sherif, "A security-enhanced robust audio steganography algorithm for image hiding using sample comparison in the discrete wavelet transform domain and RSA encryption," Multimed Tools Appl # Springer Science+Business Media New York, 2016.

[9] Lindawati and R. Siburian, "Steganography Implementation on Android Smartphone Using the LSB (Least Significant Bit) to MP3 and WAV Audio," IEEE The 3rd International Conference on Wireless and Telematics 2017.Jule 27-28, Palembang Indonesia. 2017.

[10] Sattar B. Sadkhan; Akbal O. Salman, "A survey on lightweight-cryptography status and future challenges2018 International Conference on Advance of Sustainable Engineering and its Application (ICASEA).

[11] Y.Bassil, "Audio Steganography Method for Building the Deep Web," American Journal Engineering Research (AJER) e-ISSN: 2320-0847 ISSN: 2320-0936 Volume-8, Issue-5, 2019, pp-45-51.

[12] M. Than and S. Sin, "Secure Data Transmission in MP3 file using LSB and Echo Hiding," International Journal of Advanced Research in Computer Science, 10(4), 45, 2019.

[13] Kaur N, Bansal A (2014) A review on Digital image Steganography (JCST). /International Journal of Computer Science and Information Technology 5:8135-8137.

[14] Islam AUl, Khalid F, Shah M, Khan Z, Toqeer Mahmood, et al. (2016) An improved image Steganography Technique based on MSB using Bit Differencing, IEEE 978-98.

[15] Cheddad A, Curran K, Condell J, McKevitt P, "Skin tone based Steganography in video files exploiting the YCbCr color space", IEEE International Conference on Multimedia and Expo, 2008, pages 905–908.

[16] Kousik Dasgupta, J.K.Mandal, Paramartha Dutta, "Hash Based Least Significant Bit Technique for Video Steganography(HLSB)," in International Journal of Security, Privacy and Trust Management (IJSPTM), pp 1-11, April 2012.

[17] S Khupse, N Patil, "An Adaptive Steganography Technique for Videos Using Steganoflage", International Conference on Information and Computer Technologies pages 811-815, 2014.

[18] Alaknanda S. Patil1 and Dr. G. Sundari, "Enhancing Data Security in video Steganography using Face Recognition", IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.8, August 2020, pp. 134-144.

[19] M. Coşkun, A. Uçar, Ö. Yildirim and Y. Demir, "Face recognition based on convolutional neural network," 2017 International Conference on Modern Electrical and Energy Systems (MEES), Kremenchuk, Ukraine, 2017, pp. 376-379.

[20] Deb Sunder Swami, Kandarpa Kumar Sarma, "chapter 8 A logistic-Map-Based PN Sequence for Stochastic Wireless Channels", IGI Global, 2017.

Machine Learning based Optimization Scheme for Detection of Spam and Malware Propagation in Twitter

Savita Kumari Sheoran[1]
Associate Prof. in Computer Science and Engineering Dept
Indira Gandhi University, Meerpur
Rewari, India

Partibha Yadav[2]
Ph.D. Scholar in Computer Science and Engineering Dept
Indira Gandhi University, Meerpur
Rewari, India

***Abstract*—Social networking sites are new generation of web-services providing global community of users in an online environment. Twitter is one of such popular social networks having more than 152 million daily active users making a half billions of tweets per day. Owing to its immense popularity, the accounts of legitimate Twitter users are always at a risk from spammers and hackers. Spam and Malware are the most affecting threats reported on the Twitter platform. To preserve the privacy and ensure data safety for online Twitter community, it is necessary develop a framework to safeguard their accounts from such malicious attackers. Machine Learning is recently matured and widely used technique useful to prevent the propagation of such malicious activities in social media. However, the Machine Learning based techniques have yielded a promising result in filtering the undesired contents from the user tweets but its efficiency always remains restricted within the technological limits of the technique used. To devise a more efficient model to detect propagation of spam and malware in the Twitter, this research has proposed a Machine Learning based optimization scheme based on hybrid similarity (Cosine and Jaccard) measured in conjunction with Genetic Algorithm (GA) and Artificial Neural Network (ANN). The Cosine with Jaccard in hybridization has been applied on the Twitter dataset to identify the tweets containing spam and malware. In conjunction to it the GA has been used to enhance the training rate and reduce training error by automatically selecting the designed fitness function while the ANN was applied to classify malicious tweets from through voting rule. The simulation experiments were conducted to compute the precision rate, recall, F-measures. The results of Machine Learning based optimization scheme proposed in this research were compared with the existing state-of-arts techniques already available in this regime. The comparative study reveals that the model proposed in this research is faster and more precise then the existing models.**

Keywords*—Social networking sites, Twitter, spam, malware, Cosine similarity, Jaccard similarity, genetic algorithm, artificial neural network

I. INTRODUCTION

The last two decades have witnessed an unprecedented growth in social networking sites, where people share information with the other users through radio means without verifying their identity. Several social networking sites such as Twitter, Facebook, LinkedIn, Instagram and WhatsApp, etc. have emerged as a powerful tool to facilitate the users to share information in the form of audio, video, text and pictures. These social media platforms are governed by common instinct that all of them require creating an account using personal information and need access to data computing device before actual operation. The registered users can form a socio-digital network with the other users having similar interest and can share the contents of his choice in a manner prescribed the concerned website owner. Users use these sites for varied purposes including fun, entertainments, business, and advertisement etc. For instance, Twitter, a typical micro blogging social media platform, allows users to send message up to 140 characters, make comments, attach image and pdf documents. Apart from it blogs, PDF files, picture or videos and web page can be forwarded over the platform. On twitter the registered users enjoy unrestricted access to post, like, comments, reply and re-tweet while the unregistered users can only read the tweets. Twitter users are linked in the form of an exponential hierarchy where the user's tweets are available to followers in the form of public and protected tweets.

One of astonishing feature of Twitter which differentiates it from other social media platforms is that in Twitter the relationship between users and their followers is asymmetric while in other networks it follows a symmetric or cyclic pattern. In Twitter when a user gets followers the vice-versa is not always remains true and hence followers necessarily may not have to access all the tweets of their ancestral user [1]. The tweet post is twitter can be accessed with unrestricted right by immediate followers but not to the followers at a third level of followers in tree hierarchy. But the re-tweet from second level of users will be available to third level of users. The social media communities are more liberal on their community standards and generally groups are formed between those users, who are more active and share information frequently compared to less active users. The unsolicited users enter to this chain of active users to execute their malicious activities [2]. Hackers possess as original users can have easy access to the important personal information such as bank account or passwords, available in social media account or those available in computing device (computer or mobile phone) [3]. Recent studies reveal that Twitter has become most preferred destination for cyber criminals to perform multiple malicious activities including spam, phishing and malware [4]. One of the instances of such activity was reported in March 2010 when using festive-themed messages, dangerous malware was spread

in Twitter. Later in September 2010, the malware has affected the millions of Twitter users including British Prime Minister [5-6]. Fig. 1 depicts the social- criminal ecosystem of social network especially Twitter site. As revealed in the figure, a separate community is formed by criminal users using a unique user ID along with a supporter community encircled by green dotted contour, which supports those users outside the community of criminal accounts [7]. It reveals two types of relationship among the networked community *viz.* inner and outer relationships. Inner relationship reveals the interrelation among the criminal accounts connected through social means while the outer relationship represents the interaction between the criminal accounts and his supporters, who maintain a close friendship with the criminal accounts. To propagate the threats identified in this research *viz.* malware and spam; the hackers post malicious links to unsolicited users for attracting user traffics.

The subsequent sub-sections will elaborate these both types of malicious activities for formulating further research work.

A. Spam

There is a general perception that spam mostly found in e-mails but there are instances where social networking sites frequently suffers from malicious software. Spam harms the users through various modes such as by sending undesired information in the form of advertisement or by sending messages continuously to the same e-mail or social media ID. The existing research detects the by analysing the features of the data. Beutel et al. (2013) has detected the spam by analysing the relationship between the users, social media pages and the time of instant at which the edge has been created in the social graph [8]. Another research performed by Ahmed et al. (2012) has used graph-based technique to show the relationship between the social nodes and their communication by edge of the graph [9]. The weight of edge represents the real and fake users' interactions in the form of shared URLs, pages, active friends. Here, spam detection has been performed using optimization-based machine learning approach. Sharma et al. (2014) have used Machine Learning to classify text containing spam as enunciated in the workflow [10].

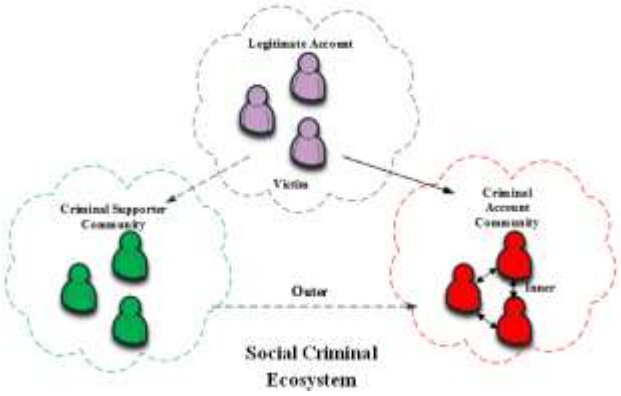


Fig. 1. Structure of the Social Criminal Ecosystem [7].

B. Malware

Malware is a type of code implanted into the network with the intention to affect the information of the legitimate users. It is injurious issue to a computer user and need to be resolved to safeguard his private and business rights. Recently there are examples where spammers have relied on outmoded social networks impersonated as e-mail to steal the information of normal users by inserting harmful worms. The Pay-Per Install (PPI) is the most amazing institute that spread malicious activities targeting the financial institutions and other websites like Facebook and Twitter. The modus operandi of Malware in affecting the information was characterized by Sanzgiri et al. (2013) [11].

This research intended to design a model to detect spam as well as malwares propagating in twitter contents through an approach based on finding similarity among the extracted features such as crime related keywords and normal keywords and the URL features. Based on the extracted features, the features are optimized using a novel fitness function of Genetic Algorithm (GA). Based on these optimized features the machine learning classifier such as ANN is trained based on the optimized features which enhance the accuracy of detection method [12]. The subsequent sections of this paper are arranged as follow: Section 2, describes the work accomplished by researcher community engaged in the regime of social network security and malicious attacks. The step by step description of the work proposed in this piece of research is presented in Section 3. The results obtained over simulation experiments and the examined parameters are discussed in Section 4. A conclusive discussion is carried out in Section 5, followed by the references consulted there in the paper.

II. Related Work

Social networking sites are naive form of socialization and hence facing out of security issues. A number of researchers have studied detection and protection of social networks against spam. Blanzieri et al. (2008) have surveyed the commonly known features which were further used to enlist the unsolicited methods of spam detection [13]. Further Sahamiet al. (1998) has deployed the unsolicited techniques such as content filtering for the same purpose [14]. In social media applications such as Twitter and Facebook the content-based methods are hardly effective because the spam contains only a few words along with the URLs. Therefore, some of the researchers have used URL blacklisting approach in order to filter the spam but this technique is not performing as per the user requirement as well as take large processing time as summarized by Grier et al. (2010) [15]. Song et al. (2011) have used relation features (distance and interconnection) among the transmitter and receiver social user for the detection of the spam in data. A list of spam and non-spam data has been created and then trained the classifier based on the extracted features. The results indicate that most of the spam has been generated by the account rather than receiver [16]. Lin et al. (2017) have presented a machine learning based approach to detect the spam based on ground truth value and provided

satisfactory performance. Also, the designed spam detection twitter model has been analysed for scalability and the performance has been measured in terms of true positive, False positive, F-score and accuracy of the system using different data size with small processing time [17].Gupta et al. (2018) have presented a spam detection framework through which the spam is identified based on user based and tweets' text-based features collectively. The use of text-based tweet features allows users to detect the spam tweet if the unsolicited user has created new account. The work has been verified based on four different classifiers such as Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF) and Gradient Boosting and Neural Network based approach has achieved highest accuracy of 91.65 among all the methods [18]. Hanif et al. (2018) have introduced additional features to measure the countermeasure in the presence of spam in Twitter site. Hanif et al. (2018) have detected the spam and malware using four machine learning techniques such as RF, SVM, *K*-Nearest Neighbour (KNN) and Multi-Layer Perceptron (MLP). They have performed a series of experiments using two simulation tools such as WEKA and RapidMiner and better results in terms of detection accuracy of 95.44% has been obtained using RF as classifier on RapidMiner tool [19]. Hai and Hwang (2018) have used deep learning as a classification approach for the detection of malware based on their malicious activities. The detection accuracy of 98.75 % has been obtained, which is quite higher as compared to the other existing techniques [20].Kaur and Sabharwal (2018) have used feed forward neural network as a classifier, which was trained based on the extracted features (+ve and -ve) in social networks. To resolve the complexity of extracted features genetic optimization has been used as an optimization approach [21].

However, a lot of researches are available in literature which studied the issue in fragmented way but the technique offering a single framework to detect spam and malware affected tweets by utilizing a minimal number of feature set is still undiscovered. To address this issue this research intends to develop a novel model to filter out the spam and malware in the Twitter using machine learning approach.

III. Proposed Work

In this research, we have applied a hybridized approach that includes GA with ANN technique to detect spammed malware. The feature of tweets has been refined and optimized based on the fitness function of GA and used dynamically to trained ANN structure. In traditional methods, the features are refined using pre-processing technique and then applied to the whole dataset. ANN is a better approach for training the data in the sense that it dynamically selects the features for the individual user data instead of applying the same features to all users. It is effective strategy for the reason that each user possesses its own characteristic features and hence need to be segregated from each other. A study on Twitter reveals that fresh accounts and Spam accounts have higher link share than that of average link shares in normal accounts. Apart from its various studies yield that spam users share more images from news web sites and roll out lucrative advertisements to lure the innocent users. Therefore, mere filtering the users sharing more images will not be sufficient to detect the spam rather filtering the users sharing more images from new websites and issuing lucrative advertisements will serve the purpose better. This study more relied on the feature of individual user group identity and classifies each user based on the URL sharing. The identified grouped in each URL's are trained using ANN approach in addition to GA scheme. Fig. 2, presents a secure framework for detecting Spam and Malware in the Twitter network.

A. Upload Data

The data used for this research were initially obtained Kaggle database and only the data related to spam, malware and normal tweets have been processed for further use in study [22]. The dataset contains total of 200K tweets along with their URL. The study was initiated on the hypothesis that all tweets contain URLs with the aim to attract social users towards malicious sites such as spam and malware downloading. The hypothesis was contradicted to obtain the URLs with spam and malware and rest were discarded.

B. Stop Word Removal

The stop word is removed by comparing each row contains in the dataset with the stored stop word list available on GitHub Gist [23]. A few stop words used in the proposed work are listed in Table I. Initially, collected data is uploaded and compared with the list of stop words in the database.

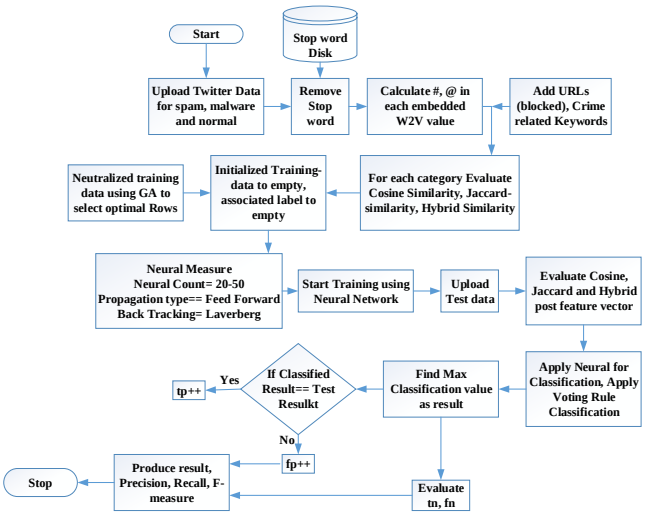


Fig. 2. Workflow Diagram of Proposed Model.

TABLE I. Stop Word List

"An"	"If"	"During"	"Before"	"After"	"Above"
"And"	"Or"	"Below"	"To"	"From"	"Up"
"But"	"Because"	"Down"	"In"	"Is"	"It"
"While"	"Until"	"Else"	"Than"	"Too"	"Very"
"Off"	"Of"	"Own"	"Can"	"Off"	"Will"
"The"	"At"	"Just"	"Don"	"Should"	"Now"

If the words in the uploaded data are matched with the database words, then these words are removed to obtain the data containing only the meaningful informative words. The mentioned algorithm is used to remove the stop word from database.

Algorithm 1: SWFD = Stop Word Removal (TUD)

Where, TUD → Twitter User Data (Individual user-wise)
SWFD → Stop Words Free Data sorted from main database

Start
Load Stop Word Dataset
Set, Count = 1
For I = 1 → All TUD
  For J = 1 → All SW
  If TUD (I, J) = SW (I, J)
  SWFD (Count) = TUD (I, J)
  Count = Count+1
  Else
  SWFD = ' '
  End – If
  End – For
End – For
Return: SWFD
End

C. Mention Ratio / URL as Content-Based Features

In this research mention ratio such as @ and # have been used as content features along with the URL. As these are the essential features used by the twitter and also used by the malicious users to misguide the normal tweet users. Therefore, it is necessary to remove these symbols from the tweet.

1) Mention Ratio: Generally, Twitter users are tagged using the '@' special character. Spammers and malware activists can also use the same special characters to trap the legitimate users. The malicious account holders entice normal users to attach with them. Equation (1) below, is used to calculate the mention ratio for each special characters.

$$Mention\ Ratio = \frac{Number\ of\ @\ present\ in\ the\ tweet}{Total\ tweets\ posted\ by\ the\ user} \quad (1)$$

2) URL Ration: Social media users generally share their thoughts and also give suggestion through tweets. The tweets posted by the sender may include URLs having a link to source pages encompassing complete information. The clever user intentionally enters a large number of URLs in their tweets to trap the legitimate users as their soft target. The URL ratio can be calculated using equation (2).

$$URL\ Ratio = \frac{Number\ of\ URL\ present\ in\ the\ tweets}{Total\ tweets\ posted\ by\ the\ Users} \quad (2)$$

3) Word to vector: The removal of stop words is followed by calculation of special characters (# and @) in the uploaded tweets, which can be applied on word to vector method. This scheme converts the text into its corresponding weighted value like as:

{-0.09450 0.16788 -0.14402 -0.0251 0.11355 -0.11794 -0.13871 -0.01607 0.1555 0.11695 0.05452 0.0936 0.08511 0.00671 -0.11653 -0.13014 0.12626 0.10248 -0.035507 -0.1523 -0.08457 0.089321 -0.01771 -0.07837 0.16123 -0.10844 -0.10118 0.03016 0.05699 0.03763 0.63156 0.06131 0.19388 -0.05652 0.1217 0.15755 0.01353 0.33352 -0.0223 -0.10877 0.11583 -0.07015 0.03653 0.05292 -0.0074 0.0242 0.08846 0.14987 0.12804 0.18679}.

The main purpose of word embedding is to study the vector representation obtained after word to vector method. One of the most commonly used word embedding method is word to vector, which maximize the probability of word condition, which is fitted in the window 'W'. After this, the crime related words appear in the URL are blocked.

On the basis of calculated value of 'W', the relationship between any two words such as W_i, W_j can be measured using hybrid similarity measures, which is a combination of Cosine Similarity and Jaccard Similarity index [24]. The similarity between tweets, which is being calculated using Cosine Similarity, is calculated using equation (3).

$$Sim_{i,j} = \frac{W_i \times W_j}{\|W_i\| \|W_j\|} \quad (3)$$

D. Cosine Similarity

Cosine similarity is a similarity analysis approach used to measure similarity score between two non-zero vectors. It is measured by the cosine of the angle between the two vectors and determines whether the two vectors point in approximately the same direction. It is often used in text analysis to measure similarity among documents. This method is used to determine the similarity and is a traditional approach, which is used in integration with the Term Frequency (TF). The text obtained after the filtering of crime related words appear in the tweet, cosine similarity is applied between the two vectors and then the multiplication of these two vectors value is being compared. Fig. 3 show the Cosine Similarity used in comparing tweets throughout this study.

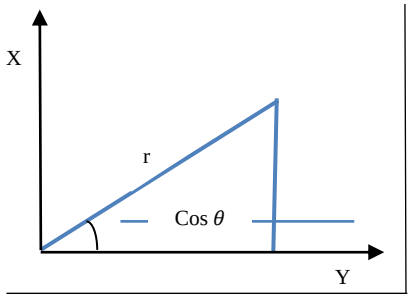


Fig. 3. Cosine Similarity.

Two tweets are declared similar if Cosine Similarity value is approaches to unity. The value of this factor approaches unity for 0^o and for other angles it is less than it [25]. The designed algorithm for Cosine Similarity is presented below:

Algorithm 2: Cosine Similarity

Required Input:	Data ← Raw data in which similarity needed
Obtained Output:	Sim_{Cos} ← Cosine similarity between data

Start
To store similarity create an empty array, Sim_{Cos}= []
Sim-count = 0
For m = 1 → Length (Data)
Current_Data = Data (m)
For n = m+1 → Length (Data)
Calculate the Cosine Similarity using given equation
L = |Cos (Current_Data) - Cos (Data (n))|
Sim_{Cos} [sim_count, 1] = Current Data
Sim_{Cos} [sim_count, 2]= Data(n)
Sim_{Cos} [sim_count, 3]=L
Incremental array →Sim-count = Sim-count + 1
End – For
End – For
Return: Sim_{Cos} as final output of cosine similarity between data
End – Function

E. Jaccard Similarity

Jaccard similarity is used to determine the similarity as well as the distinction among the documents based upon the attributes. Its value lies between 0 to 100 percentages. Higher percentage value represent more similar is the data while lower value infers least similarity. An effort has also made to determine the similarity using Jaccard Similarity with relationship between two tweets by calculating Jaccard Coefficient, basically utilized to compare data based on similarity, dissimilarity and distance bases [26]. The output obtained using Jaccard similarity is the rate of number of tweet features that are most common to the entire text with respect to the number of features present in the entire tweet. The measured similarity calculated using Jaccard similarity is given by equation (4).

$$J\left(W_i, W_j\right) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|} \quad (4)$$

Following algorithm is implemented for Jaccard Similarity:

Algorithm 3: Jaccard Similarity

Required Input:	Data ← Raw data in which similarity needed
Obtained Output:	Sim_{Jac} ← Cosine similarity between data

Start
Create an empty array to store similarity, Sim_{Jac} = []
Sim-count = 0
For m = 1 → Length (Data)
Current_Data = Data (m)
For n = m+1 → Length (Data)
Union = (Cos (Current_Data) U Cos (Data (n)))
Intersection = (Cos (Current_Data) ∩ Cos (Data (n)))
Sim_{Jac} (sim_count) = $\frac{Count\ (Union)}{Count\ (Intersection)}$
Incremental array →Sim-count = Sim-count + 1
End – For
End – For
Return:Sim_{Jac} as Jaccard Similarity between data
End – Function

F. Genetic Algorithm (GA)

GA has been used as a feature selection algorithm in order to select the row features of the tweets obtained after hybridizing Cosine and Jaccard Similarity Index. Feature selection is one of the essential tasks, that helps to enhance the training accuracy of the classification algorithm such as Neural network is used to train the system based upon the optimized features obtained as per the designed fitness function as denoted by equation (5).

$$F(f) = \begin{cases} 1\ (Selected)\ if\ (1 - f_s);\ Max\ among\ all\ vectors < f_t \\ 0\ (Not\ Selected)\ Otherwise \end{cases} \quad (5)$$

Where,

F_t Generated mutation error

f_s : Current feature in FD

f_t: Threshold feature and it is the average of all FD

The implementation in GA can be accomplished in three steps depicted in the Fig. 4:

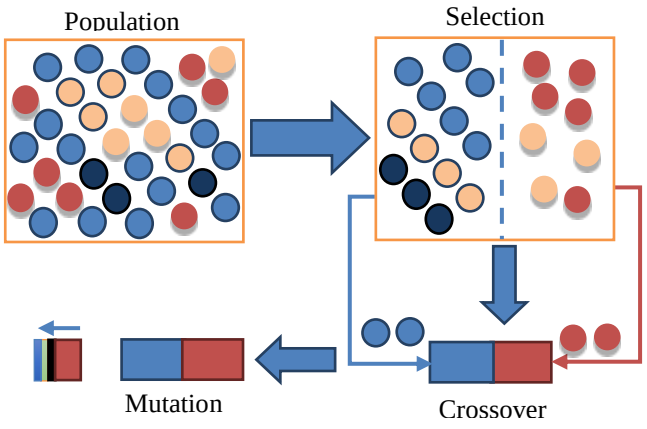


Fig. 4. Steps in GA Implementation Process.

GA is a basic heuristic algorithm that works on Darwin's theory of evolution and is also named as Evolutionary Algorithm that finds the best solution based on the natural selection and crossover as shown in Fig. 4. Genetic algorithms randomly generate a set of populations. A distinct gene is comprised by each individual and hence responsible for different solution to a particular problem, which is again encoded by the chromosomes. To solve the problem, a problem specific objective function is designed. GA mainly included three operators *viz.* (i) Selection (ii) Crossover and (iii) Mutation. Selection is used to choose individuals from the present generation, which is later used for next generation. At this stage the best one with high fitness values are selected. In chromosomes, it is responsible to suggest parents (two best chromosomes that are responsible for best generation). This process is repeated until the desired solution is obtained. The workflow of GA is written in algorithmic form as below:

Algorithm 4: Features Selection using GA

Required Input:	Feature Data ← Extracted feature from used Dataset
	Fitness Function ←Designed fitness function for feature selection
Obtained Output:	OFD←Optimized Feature Data

Start Feature Selection
Load Dataset, Feature Data (FD) = Load feature sets
To optimized the FD, Genetic Algorithm (GA) is used
Set up basic operators and parameters of GA:
Population Size (P) – Based on the number of properties
CO – Crossover Operators
MO – Mutation Operators
OFD – Optimized Feature Data
Calculate fitness function [F(f)]with usual terms

$$\mathbf{F(f)} = \begin{cases} \mathbf{1}\ (Selected) if (\mathbf{1} - \mathbf{f_s});\ Maxamongallvectors < f_t \\ \mathbf{0}\ (NotSelected) Otherwise \end{cases}$$

Set,Optimized Feature Data, OFD = []
For i in rang of R
F_s = FD (i) = $Selected_{Feature}$
$F_t = Threshold_{Feature} = \sum_{i=1}^{R} FD(i)$
$F(f) = Fit\ Fun\ (F_s, F_t)$
Nvar = Number of variables
$Best_{Prop}$ = OFD = GA (F(f), T, Nvar, Set up of GA)
End - For
Return: OFD as an Optimized Feature Data
End – Function

G. Artificial Neural Network (ANN)

After optimizing the features based on the fitness function of GA as according to equation (5), these features are used to train Neural Network as a classification algorithm. ANN is designed to work in the same way as that of human brain. Its working is inspired by the biological nature of cell known as Neurons or sometimes knows as nodes. The structure of ANN with 'N' number of data input and single output is shown in Fig. 5 while Fig. 6 shows the examined Mean Square Error (MSE) value during the training process of a spam and malware detection based social media system.

The figure shows that the desired value has been obtained after passing the 20 number of neurons to the hidden layer of ANN structure.

Fig. 5. Trained ANN Structure with MSE Value.

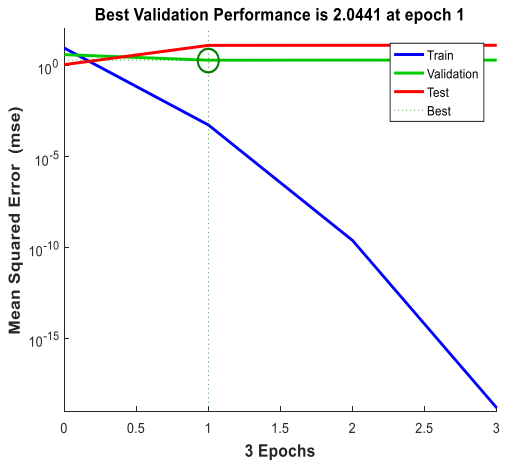


Fig. 6. Mean Square Error for Epoch.

Also, The MSE value examined during the training process ID indicated by the blue line, which is approaches to zero at 3rd iteration. The designed algorithm for ANN is represented as follows:

Algorithm 5: Training using ANN

Required Input:	OFD←Training Data as an optimized feature data
	C←Target/Category in terms of spam, malware and normal data
	N←Number of Neurons
Obtained Output:	Net ← Trained structure

Start Detection
Load Training Data, T-Data **= OFD**
Declare the initial parameters of ANN
- Epochs Counts: E
- Neurons Counts : N
- Performance Parameters: MSE, Gradient, Mutation and Validation
- Techniques Applied: Levenberg Marquardt Algorithm
- Data Division Strategy: Random

For i = 1 → T-Data
If T belongs to spam
Group (1) = Features (OFD)
Else if T belongs to malware
Group (2) = Features (OFD)
Else // Normal Case
Group (3) = Features (OFD)
End – If
End – For
Implement the ANN through Training data and Group
Net = Newff $(T - Data, Group, N)$
Setting training parameters as per the requirements and accomplish the train task
Net = Train (Net, T-Data, Group)
Return: Net value according to trained structure
End – Function

The testing of spam and malware detection social system has been performed by uploading the tweets as test data and then measure the similarity among the uploaded documents using Cosine with Jaccard as similarity measure. The data obtained are compared with the data stored into the ANN database by applying the voting rule as a cross-validation scheme. Here the voting classifier is used in addition to ANN classifier. If maximum value has been obtained, then calculate True Negative (T_n) and False Negative (F_n) values for the uploaded data. In case, if classified results are equal to test results then True Positive (T_p) and False Positive (F_p) has been calculated. Subsequent section 4 of this paper presents and discuss the results obtained for the parameters (T_n), (F_n), (T_p) and (F_p) in term of precision, recall, F-measure.

IV. Results and Discussion

The performance analysis of proposed model was carried out through simulation experiments conducted using standard settings considering optimization, classification with similarity measurement tools. A total of N-700 tweeter data were analysed over Simulink and Natural Language toolkit for parametric analysis and stop word removal respectively. The performance has been measured for three parameters precision, recall, F-measure using standard equations (6), (7) and (8) respectively. Where Precession signify the instances of correctness in the experiment, Recall signify the measure of correct hit and F-measure score is related to accuracy or correct prediction per unit of input.

$$Precision = \frac{T_p}{T_p + F_p} \quad (6)$$

$$Recall = \frac{T_p}{T_p + F_n} \quad (7)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

Where

T_p = Number of tweets that are actually spam/ malwares and also predicted as malicious.

F_n = Number of tweets that are being predicted as real but are spam and contains malwares.

F_p = Number of tweets that is actually real but predicted as affected one (Spam/ malwares).

T_n = Number of appropriately predicted real tweets.

The variation of precision values with number of tweets uploaded for various techniques *viz.* Cosine, Jaccard, Hybrid and GA with ANN approach are presented in Fig. 7. Figure illustrates that proposed work implementing GA with ANN in combination with hybrid similarity measure have highest Precession values for any number of tweets. The average precision computed for cosine, Jaccard, hybrid and GA with ANN approach are 0.746, 0.805, 0.885 and 0.963 respectively which reveals that the tweet that are filtered as a sub part of spam or malware for the tested dataset is maximum for GA with ANN approach (proposed in this research).

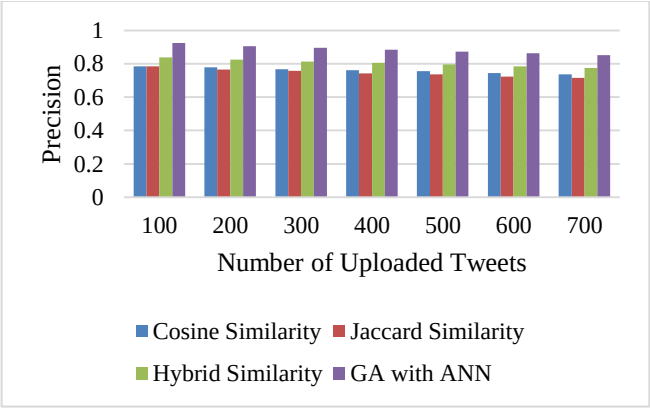


Fig. 7. Precision versus Number of uploaded Tweets (N=700).

The recall parameter represents the rate of tweets that are being posted by genuine user and have been predicted as spam or malware by the user accurately. The examined value of Recall for the uploaded tweet in the range from 100 to 700 is shown in Fig. 8. The average recall rate examined for the Cosine similarity, Jaccard Similarity, hybrid similarity and GA with ANN are 0.694, 0.785, 0.864, and 0.894 respectively which reveals that the tweet that are filtered owing to genuineness against spam or malware for the tested dataset is maximum for GA with ANN approach (proposed in this research).

To represent the arithmetic means of precision and recall of the examined values F-measure is illustrated in Fig. 9. F-measure basically envisages the accuracy of a model. The examined average values of F-measure for four different schemes *viz.* Cosine similarity, Jaccard Similarity, hybrid similarity and GA with ANN are 0.719, 0.822, 0.874, and 0.927 respectively which again reveals that the tweet that are filtered per unit of input due to spam or malware for the tested dataset is maximum for GA with ANN approach (proposed in this research).

A comparison of average values of the parameters under study for proposed model with the existing state-of arts in the area of research viz. K. Subba and E. Srinivasa (2019) [27] and Murugan and Devi (2018) [28] is tabulated in Table II and represented graphically in Fig. 10.

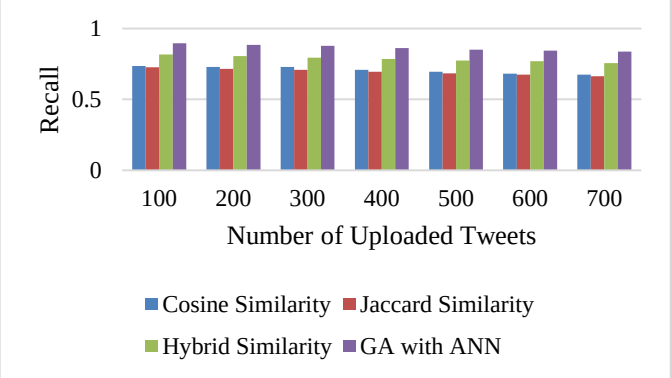


Fig. 8. Recall versus Number of uploaded Tweets (N=700).

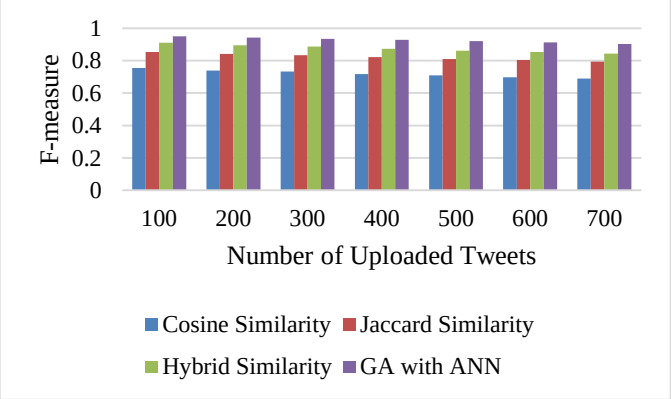


Fig. 9. F-Measure versus Number of uploaded Tweets (N=700).

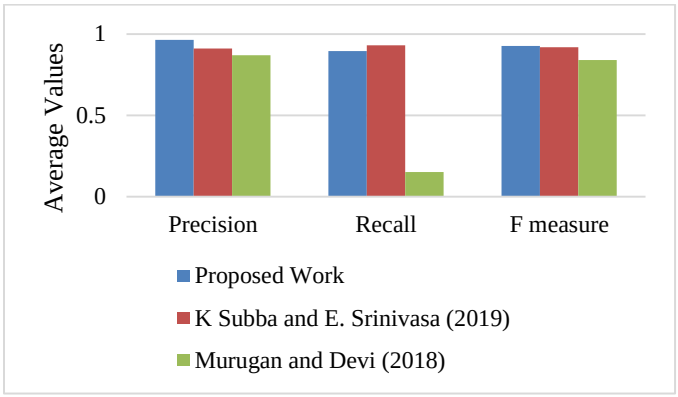


Fig. 10. Comparison of uploaded Tweets (N=700).

TABLE II. COMPARISON OF PARAMETERS

Parameters	Proposed Model	K. Subba and E. Srinivasa (2019) [27]	Murugan and Devi (2018) [28]
Precision	0.963	0.91	0.87
Recall	0.898	0.93	0.15
F-measure	0.927	0.919	084

A detailed look at the available literature reveals that the models established by K. Subba and E. Srinivasa (2019) [25] and Murugan and Devi (2018) [28] are state-of-arts exiting models having best performance so far. The comparison of performance of the existing state-of-arts with the model proposed in this research is shown in Fig. 10. Above results reveals that model proposed by us using the hybrid GA with ANN approach outperform the Murugan and Devi (2018) on all three examined parameter while it outperform the K. Subba and E. Srinivasa (2019) Model on the two parameters *viz.* Precision and F-measure and almost lessen Recall value. For quantitative purpose the precession in filtering the spam and malware for the proposed model is improved by 5.82 % and 10.69 % respectively from K. Subba and E. Srinivasa (2019) [25] and Murugan and Devi (2018) [28] models. Therefore, overall performance of the model proposed in this research is better than the existing models.

V. CONCLUSION

Presently Social networking sites are the most popular mode of network formation for the purpose of exchanging the information, advertise and the business purpose. Owing to their global popularity the Social Networking sites are at a great risk of having been used to misguide the genuine users from malicious activities of spammers and malwares. Therefore, to ensuring the data safety and privacy of the social media user is a need hour. In literature the measurement of Cosine Similarity, Jaccard Similarly and Hybrid Similarity has been carried out to evaluate the Precession, Recall and F-measure values for decide the effectiveness of a model in preventing the spam and malware but improving the performance is always remained an open challenge before research community working in this regime In this paper, we have designed a secure threat prevention (spam and malware) system for Twitter site.

We have used Machine Learning approaches such as GA and ANN in hybridization of existing models involving measurement of Cosine and Jaccard similarity. In our model novel GA approach and has been used for classification and ANN with voting algorithm is used for cross validation purpose. The simulation study carries out on N=700 tweets, reveals that average precision, recall and f-measure of 0.963, 0.894 and 0.927 has been achieved which is 5.82 % and 10.69 % higher than the other two models *viz.* K. Subba and E. Srinivasa (2019) and Murugan and Devi (2018); used as standard reference in research. This study reveals that the Machine Learning is an effective tool for prevention of legitimate users against attack of spam and malwares. Here in this research we have applied GA and MLL for filtering some stop words from Twitter and observed a promising result. In future we are planning to further investigate the similar issues using deep learning approach with complete text analysis with NLP in Twitter and social media sites. Such study may yield more effective results for preventing malicious attack of legitimate social media users.

REFERENCES

[1] A. Sanzgiri, A. Hughes, and S. Upadhyaya, "Analysis of malware propagation in Twitter," IEEE 32nd International Symposium on Reliable Distributed Systems. pp. 195-204, 2013.

[2] G. Fei, H. Li, and B. Liu., "Opinion Spam Detection in Social Networks," In Sentiment Analysis in Social Network, pp. 141-156, 2017.

[3] D. Niranjan Koggalahewa, Y. Xu, and E. Foo, "Spam Detection in Social Networks based on Peer Acceptance," In Proceedings of the Australasian Computer Science Week Multiconference, pp. 1-7, February 2020.

[4] B. A. Kamoru, A. Jaafar, M. B. Jabar, and M. A. Murad, "A mapping study to investigate spam detection on social networks," Int J Appl Inform Syst, vol. 11, no. 11, pp. 16-31, 2017.

[5] M. R. Faghani and H. Saidi, "Malware propagation in Online Social Networks," IEEE 4th International Conference on Malicious and Unwanted Software (MALWARE,) pp. 8-14, 2009.

[6] K. Nagaramani, K. Vandanarao, and B. Mamatha, "Machine Learning Algorithms for Spam Detection in Social Networks," Asian Journal of Computer Science and Technology, 8(S3), pp. 41-44, 2019.

[7] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," In Proceedings of the 21st international conference on World Wide Web, pp. 71-8, April 2012.

[8] A. Beutel, W. Xu, V, Guruswami, C. Palow and C. Faloutsos, "Copycatch: stopping group attacks by spotting lockstep behavior in social networks," In Proceedings of the 22nd international conference on World Wide Web, pp. 119-130, May 2013.

[9] F. Ahmed and M. Abulaish, "An MCL-based approach for spam profile detection in online social networks," In 2012 IEEE 11th international conference on trust, security and privacy in computing and communications, pp. 602-608, June 2012.

[10] N. Sharma,and A.Verma, "Survey on Text Classification (Spam) Using Machine Learning," (IJCSIT) International Journal of Computer Science and Information Technologies, 5(4), pp. 5098-5102, 2014.

[11] A.Sanzgiri, A. Hughes, and S. Upadhyaya, "Analysis of malware propagation in Twitter," IEEE 32nd International Symposium on Reliable Distributed Systems, pp. 195-204, September 2013.

[12] F.J. Alqatawna, A. Madain, Z.A Ala'M and R. Al-Sayyed, " Online social networks security: Threats, attacks, and future directions," In Social Media Shaping e-Publishing and Academia, pp. 121-132, Springer, Cham, 2017.

[13] E. Blanzieri, and A. Bryl, "A survey of learning-based techniques of email spam filtering,"Artificial Intelligence Review, 29(1), pp. 63-92, 2008.

[14] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz,"A Bayesian approach to filtering junk e-mail," In Learning for Text Categorization Papers from the 1998 workshop,Vol. 62, pp. 98-105, July 1998.

[15] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," In Proceedings of the 17th ACM conference on Computer and communications security, pp. 27-37, October 2010.

[16] J. Song, S. Lee, and J. Kim, "Spam filtering in twitter using sender-receiver relationship," In International workshop on recent advances in intrusion detection, pp. 301-317,Springer, Berlin, Heidelberg, September 2011.

[17] G. Lin, N. Sun, S. Nepal, J. Zhang, Y. Xiang, and H. Hassan, 'Statistical twitter spam detection demystified: performance, stability and scalability," IEEE access, 5, pp. 11142-11154, 2017.

[18] H.Gupta, S.M. Jamal, S. Madisetty, and S.M. Desarkar, "A framework for real-time spam detection in Twitter," 10th International Conference on Communication Systems & Networks (COMSNETS), pp. 380-383, IEEE, January 2018.

[19] M. H. M. Hanif, S. K. Adewole, B. N. Anuar, and A. Kamsin, "Performance Evaluation of Machine Learning Algorithms for Spam Profile Detection on Twitter Using WEKA and RapidMiner," Advanced Science Letters, 24(2), pp. 1043-1046, 2018.

[20] T.Q. Hai, and O.S. Hwang, "An efficient classification of malware behavior using deep neural network," Journal of Intelligent & Fuzzy Systems, 35(6), pp. 5801-5814, 2018.

[21] J. Kaur, and M. Sabharwal, "Spam detection in online social networks using feed forward neural network," In RSRI conference on recent trends in science and engineering 2, pp. 69-78, 2018.

[22] https://www.kaggle.com/uciml/sms-spam-collection-dataset accessed on January 01, 2020.

[23] https://gist.github.com/sebleier/554280 accessed on January 02, 2020.

[24] X. Yang, C. Macdonald, and I.. Ounis, "Using word embeddings in twitter election classification," Information Retrieval Journal, 21(2-3), pp. 183-207, 2018.

[25] R.A. Lahitani, E.A. Permanasari, and A.N. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," 4th International Conference on Cyber and IT Service Managemen, pp. 1-6. IEEE, April 2016.

[26] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard coefficient for keywords similarity," In Proceedings of the international multiconference of engineers and computer scientists 1 (6), pp. 380-384, March 2013.

[27] K Subba Reddy, E. Srinivasa Reddy, " Using Reduced Set of Features to Detect Spam in Twitter Data with Decision Tree and KNN Classifier Algorithms," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, 8 (9), pp 6-12, 2019.

[28] S.N. Murugan, and U.G Devi, "Detecting streaming of Twitter spam using hybrid method," Wireless Personal Communications, 103(2), pp. 1353-1374, 2018.

Secure Intruder Information Sharing in Wireless Sensor Network for Attack Resilient Routing

Venkateswara Rao M[1]
Research Scholar
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundaion
Vaddeswaram,Andhra Pradesh, India

Srinivas Malladi[2]
Professor
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation,
Vaddeswaram,Andhra Pradesh, India

***Abstract*—Securing the routing process against attacks in wireless sensor network (WSN) is a vital factor to ensure the reliability of the network. In the existing system, a secure attack resilient routing for WSN using zone-based topology is proposed against message drops, message tampering and flooding attacks. The secure attack resilient routing provides a protection against attacks by skipping the routing towards less secure zones. However, the existing work did not consider the detection and isolation of the malicious nodes in the zone based wireless sensor network. To solve this issue, we proposed enhanced attack resilient routing by detecting malicious zones and isolating the malicious nodes. We proposed a three-tire framework by adopting sequential probability test to detect and isolate malicious nodes. Attacker information is shared in a secure manner in the network, so that routing selection decision can be made locally in addition to attack resiliency route selection provided at the sink. Overhearing rate is calculated for all nodes in each zone to detect blackhole attackers. Simulation results shows that the proposed Three Tier Frame work provides more security, reduced network overhead and improved Packet delivery ratio in WSNs by comparing with the existing works.**

Keywords*—Flooding; malicious zone; network overhead; overhearing rate; packet delivery ratio

I. INTRODUCTION

Wireless sensor network (WSN) technology is growing rapidly in many emerging sectors such as industry monitoring and control, home surveillance, wild life monitoring and smart farming. WSN is a network created by sensors equipped with wireless transceivers for communications. Sensor nodes collect environment parameters and send it to a central station for processing. The sensor node can send data to the central station in one hop or via multi hop forwarding depending on the distance between sensor node and the sink. Due to unattended nature and wireless infrastructure, the sensor network is easily susceptible to various kinds of attacks like message dropping, message tampering, message flooding etc. These attacks must be detected and mitigated to ensure the reliability of the sensor networks.

In [1] a secure attack resilient routing protocol is proposed to secure the routing process against possible attacks such as message dropping, message tampering and flooding. The whole network is divided into several zones and each zone is scored on the basis of the security and energy available in the zone. The zone with low security score is not preferred by the sink node for routing the packets. But this work did not specifically identify the malicious nature of the zone and did not isolate the specific attacker node. All zones are scored based on security and energy availability in that zone. Due to energy imbalance in the zones, still there is a higher chance of selecting less secure route by sink node.

It is important to identify the malicious zone and isolate the specific attacker node in that zone for secure data transmission. The attacker or compromised node may fabricate or tamper the data packet in the routing and it is a major problem to solve.

To solve this problem, a three-tier framework is proposed for secure attack resilient routing to transmit packets in secured manner from source node to destination node. The attacker node is detected and the information about the attacker node is shared in a secure manner in the network. Through sharing of attacker information, the network is made attack resilient and other innocent nodes are aware about the attacker node. Hence packets are routed only through innocent nodes in a secured manner.

Watch dog mechanism is employed by monitoring node in three tier frame work to detect the attackers. Monitoring node runs in promiscuous mode and observes all the packets within the zone. The monitoring node calculate overhearing rate (OR) for all nodes within its zone. Black hole attackers are identified Based on overhearing rate calculated by monitoring node. Sequential probability hypothesis test is used to check whether the Node is selective dropper or not. Monitoring node observes the rate of packets generated by the nodes in the zone and when rate exceeds certain threshold, it detects the node as flooding node.

Monitoring node shares the information about the message dropping and message tampering attackers to the sink in a secure way, so that sink can skip these routes while processing the packet for routing.

If monitoring node observes flooding attack from a node in a zone, it generates a blacklist packet containing the flooding node information. The packets from the blacklisted node are dropped by the nodes in the zone. Therefore, the effect of the flooding attack is restricted as much as possible.

The remaining part of the paper is organized as: The review of related work is presented in Section II. The

Proposed solution is elaborated in Section III. Detailed results are discussed with the help of tables and charts in Section IV. Conclusion and further enhancements for proposed work are depicted in Section V.

II. Related Work

Compromised nodes are detected using statistical analysis in [2]. Based on the past observations, sink calculates the probability for a node to be malicious. The overhead of detection is at sink. In [3], a light weight defense mechanism against black hole attack is proposed. Based on observation of packet sequence number, message droppers are identified. Once black hole message droppers are identified, ICMP control packets with information of black hole attackers is broadcasted in the network. So black hole nodes are skipped in the routing. Authors in [4] proposed a method to detect and alleviate from cooperative black hole attack. The detection of black hole is based on absence of consistent acknowledgement for the packets. The black hole node can be precisely located in this solution. The sensitive regions where there is high probability of packet loss are identified and routing through these paths is prevented using a sensitive guard procedure. E-watch dog mechanism is proposed in [5] to detect selective message droppers. This scheme is proposed to solve the problem of higher false positives in the traditional watch dog mechanism of packet monitoring. To solve the higher false positives in watch dog monitoring, the placement of monitoring hidden node problem is avoided. Time required for attacker detection, False positives, Network overhead and Accuracy of detection is measured for performance analysis. A heuristic solution for attack detection is proposed in [6]. In this work attackers are detected at the route discovery stage by observing the discrepancy in the sequence number of route request and route reply. Due to detection at route discovery stage, overhead for attack detection is less in this methodology. Black hole nodes are detected using cooperative sensing in [7]. A semi centric detection process called BlackDP is proposed in [8]. The solution can detect cooperative black hole nodes and isolate them in a two-stage process. In first stage any suspicious activity in the route reply with highest sequence number is notified to a cluster head. In the second stage, cluster head verifies all suspicious nodes and shares the information about blacklist to all nodes within the cluster to proactively drop the blacklisted nodes in the routing path. Authors in [9] proposed a solution to secure against cooperative black hole nodes in the MANET. Designated monitoring nodes are called security monitoring nodes and they are deployed in certain places in the network. Monitoring nodes detect black hole attackers by probing the packets and on detection of attack, the information is shared periodically to rest of the nodes. A cross layer protocol for detecting cooperative black hole nodes is proposed in [10]. The solution is based on watch dog monitoring of RTS/CTS at the MAC layer and to solve the problem of false alarm in watch dog monitoring, which is done by network layer. In [11], AODV protocol is extended for detecting multiple black hole attacks in the network. The black hole nodes are detected by monitoring the discrepancy in the count of packets. The node that detects the attacker, shares the attacker information to rest of the nodes in the network. The nodes maintain a dynamic blacklist to keep the information of black list nodes and proactively skip those black hole attackers from the routing path. A light weight black hole attack detection method is proposed in [12]. Cluster heads are deployed redundantly. Passive cluster heads use watchdog monitoring mechanism to detect compromised cluster heads. Authors in [13] proposed a black hole attack detection using acknowledgement scheme. Special designated nodes called monitor nodes are deployed in the network. Destination node sends an acknowledgement for each packet received from source node. This acknowledgement is monitored by monitoring node to detect packet loss due collisions. The traditional AODV protocol is integrated with bait detection scheme to detect collision attacks. On detection of black hole nodes, monitoring node forwards the information to rest of nodes to prevent blacklisted nodes from routing packets. Packet delivery ratio, Time required for attacker detection, False positives, Network overhead and Accuracy of detection is measured for performance analysis. Hidden Markov Model is applied for message drop attack detection in [14]. The nodes in the relay path are analyzed using Hidden Markov Model to detect the message drop attacks. Information about malicious nodes is sent to all other nodes in the network to mitigate the impact of such nodes in the routing path. A centralized geo-statistical hazard model to detect malicious regions in the network is proposed in [15]. Detection and mitigation of attacks is not handled uniformly in the network. Base station samples, analyze the suitability of the area for detection and launches detection only in the selected areas. A group-based technique for detection of multiple message drop attacker in the network is proposed in [16]. The clustering topology is used solution. The detection of message drop attack is done by the cluster head nodes. Cluster head nodes send probing messages to the nodes in the cluster and wait for acknowledgements. Based on acknowledgement monitoring, message droppers are detected and isolated in the network. Authors in [17] analyzed the recent trends in security of wireless sensor networks. Authors in [18] proposed analytical model for analyzing the security of wireless communications. Work in [19] and [20] identified malicious nodes and isolated them using certificate revocation. In addition to end-to-end delay, the propagation of large amount of data in MANETs is liable for higher energy usage, thereby influencing the parameters such as network efficiency, throughput, packet overhead, energy usage. To increase the longevity of the network and energy usage, efficient parameter metric measures are adopted in [21], [22].

III. Proposed Solution

A. Network Model

Each sensor node in WSN contains a unique ID. It is preconfigured with the private key of Hyperelliptic curve cryptography (HECC) and the corresponding public key is maintained at sink node. Hyperelliptic curve is a type of elliptic curves with genus ≥ 1. Elliptic curve cryptography (ECC) is found to have lower complexity than RSA. But still the complexity is high in ECC considering the case of resource constrained wireless sensor network. HECC is proposed to solve this problem. Equation (1) represents Hyperelliptic curve C with genus g over k.

$$C: y^2 + a(x)y = b(x) \quad (1)$$

Where

a(x): A polynomial with degree ≤ g over b

b(x): A monic polynomial with degree 2g+1 over b

Equation (2) is an example for sample HECC Function

$C: y^2 = x^5 - 5x^3 - 4x - 1$ over Q genus g=2. (2)

The public and private key pair of source node and sink is unique and not available to other nodes Also, the secret key sequence and a hash function H is assigned to each node in WSN. The secret key sequence and H is known to the source node and sink.

The whole WSN is split into $M \times M$ zones. The zone size is set in such a way that nodes in the same zone are one hop away from each other. For each zone, a node close to the center of the zone is selected as the monitoring node.

B. Secure Intruder Information Sharing

The architecture of secure intruder information sharing in WSN shown in Fig. 1.

A three tier frameworks is proposed to solve the secure intruder information sharing problem in wireless sensor networks.

- At the top tier is sink node, which prevents routing to risk zones and blocks the transition of Route Reply (RREP) containing risk zone relays.
- At the middle tier is the monitoring nodes [13]. They do not participate directly in routing, but instead passively monitor the packets and detect attacks within the zone and share this information to neighboring zones and sink.
- At the bottom tier is the ordinary sensor nodes and they send data through multi hop routing to the sink node.

There are two functionalities in the three-tier framework

- Detection of attack.
- Mitigation of attack.

1) Detection of attack: Watch dog mechanism is employed by the monitoring node to detect the attacks. Monitoring node runs in promiscuous mode and observes all the packets within the zone. The monitoring node calculate overhearing rate (OR) for all the nodes within its zone. The OR value is calculated by observing the RTS/CTS packets in the MAC layer using the Equation (3).

$$OR = \frac{TF}{TO} \quad (3)$$

Where TO is the count of overheard packet and TF is the count of forwarded packets? Every time monitoring node overhears packet TO is incremented and whenever monitoring node finds the overheard packet is forwarded TF is incremented. When the overhearing rate is continuously less than a threshold value, the corresponding node can be confirmed as black hole attacker.

But deciding on selective message dropper cannot be made based on OR threshold alone and monitoring node relies on sequential probability test to confirm the selective message dropper in this work. Sequential probability test is a statistical testing technique to check the validity of a hypothesis based on observation over a period of time.

Sequential probability test tries to prove one of the following hypotheses.

H_0: Node is not a selective dropper.

H_1: Node is a selective dropper.

To prove the hypothesis this work uses two thresholds A (upper) and B (lower) based on false positive rate α and false negative rate β [5] as shown in Equations (4) and (5).

$$A = \log \frac{\beta}{1-\alpha} \quad (4)$$

$$B = \log \frac{1-\beta}{\alpha} \quad (5)$$

The tolerant value for α, β is set by the monitoring node.

The log probability for a node x for T tests is given in

Equation (6)

$$P(x) = \log \frac{\prod_{t=1}^{T} P_1(S_t)}{\prod_{t=1}^{T} P_0(S_t)} \quad (6)$$

The following observations can be done Based on P(x).

Hypothesis H0 can be accepted if P(x)<A and the test can be stopped for the node x.

Hypothesis H1 can be accepted if P(x)>B and the test can be stopped for the node x. In this case, monitoring node marks the node as selective dropper.

For A<P(x)<B, both of the hypothesis cannot be confirmed now and further test is needed for node monitoring node observes the rate of packets generated by the nodes in the zone and when rate exceeds certain threshold, it detects the node as flooding node.

Monitoring node correlates the received and forwarded packets and checks if the packet content is altered. On detection of alteration of packets, the node which is forwarding can be identified as message tampering attacker.

Monitor node is able to detect message dropping, message tampering and message flooding attackers through the process of promiscuous monitoring.

2) Mitigation of attack: Monitoring node shares the information about the message dropping and message tampering attackers to the sink in a secure way, so that sink can skip these routes while processing the route reply (RREP).

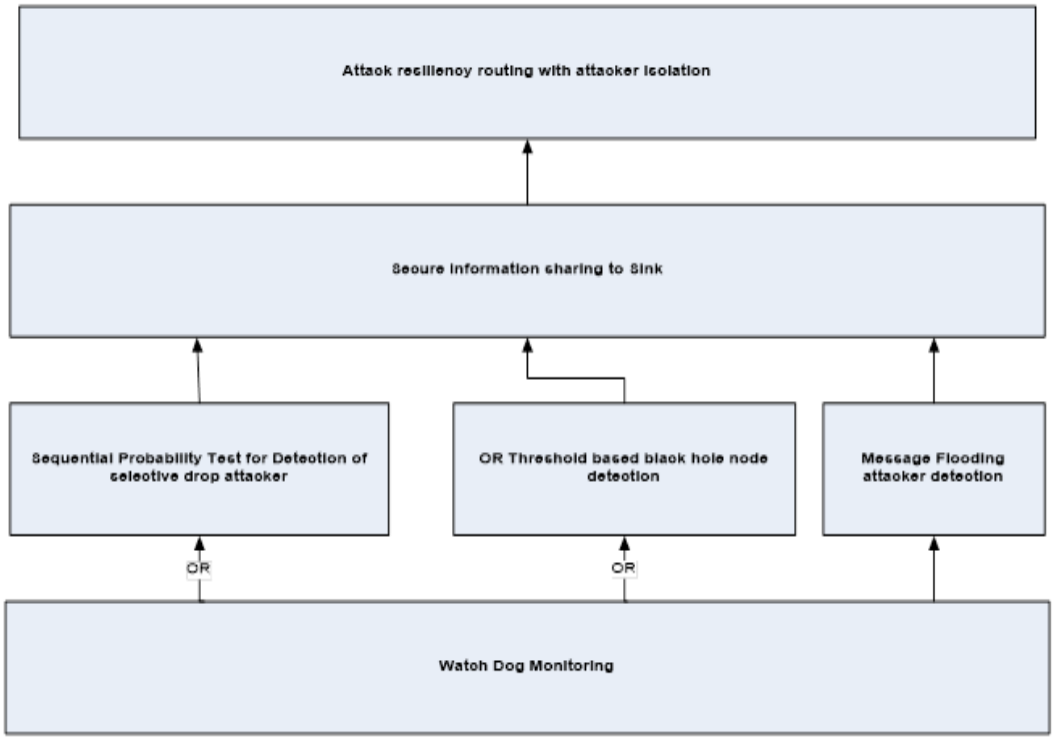


Fig. 1. Architecture of Secure Intruder Information Sharing.

Monitoring node sends the information of message dropper and message tampering attacker to sink node via a new packet type called blacklist. The blacklist packet has following format:

BlackList
{
Source
Timestamp
Encrypted payload
}

The encrypted payload has information of the attacker found. The encryption is done using HECC private key and sent to the sink. Encryption process is shown in Fig. 2.

If the packet is dropped in the zone, the monitoring node observes it and then attempts the cooperative forward for relaying the packet. Cooperative forwarding mechanism ensures the reliability of the BlackList packet.

Once the BlackList packet is received at sink, it decrypts the Encrypted payload using the HECC public key. The nodes found after decryption is added to a blacklist maintained at the sink. The Decryption process is shown in Fig. 3.

When RREQ is received at sink, before processing it for sending RREP sink checks the nodes in the RREQ for their presence in the blacklist maintained at the sink. In case of presence, RREP is not generated for the paths. Only from the rest of the paths, the one with highest security score is selected and RREP is generated with that path.

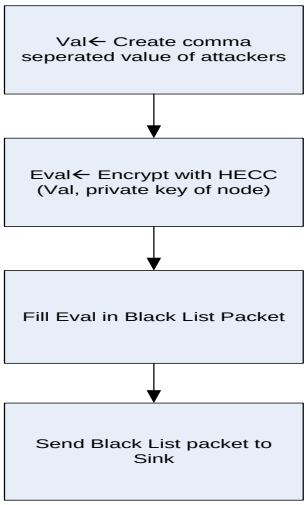


Fig. 2. Encryption Process.

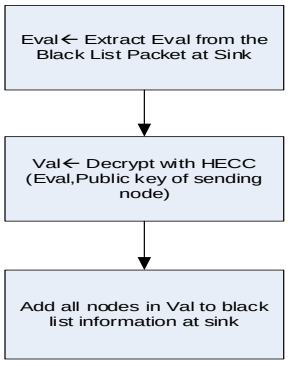


Fig. 3. Decryption Process.

In case monitoring node observes flooding attack from a node in a zone, monitoring node generates an ALERT packet containing the flooding node information. The ALERT packet is broadcasted to next immediate neighboring zones. The neighboring zone nodes and local zone nodes, after receiving ALERT packet, add the node in the ALERT packet to their blacklist. The packets from the blacklisted node are dropped by the nodes in the zone. Therefore, the effect of the flooding attack is restricted as local as possible.

3) Novelty in proposed solution: The proposed solution has better performance than the solutions reported in earlier works [5] and [13] in the following ways:

- Detection effort is localized within zone, thus reducing the unnecessary overhead.
- Mitigation is distributed in both sink and neighboring.
- Zones, thereby there is a more control on the attackers.
- Sharing of information between the zone and the sink is secured using HECC algorithm, thereby it is difficult to tamper the information about the attacker.
- A highly reliability for packet carrying attacker information is ensured.

IV. Results

Simulation was conducted in NS2 for proposed solution with the parameters shown in Table I.

The solution for proposed work is compared with solution proposed in [5] for selective attacker detection and solution proposed in [13] for detection malicious attacker in sensor network.

In terms of the following parameters, the performance of the proposed and existing works is compared.

- Packet delivery ratio
- Accuracy of detection
- False positives
- Time for attack detection
- Network Overhead

The ratio of number of packets received at sink to the number of packets sent from source to sink is termed as packet delivery ratio. The rationale for measuring the packet delivery ratio is to measure the resilience of the packet transmission in the network in presence of message dropping attacks.

The packet delivery ratio is calculated by varying the number of nodes with 10% of nodes as attackers and the result is presented Table II and plotted Fig. 4.

The packet delivery ratio in the proposed work is 7.65% more than that of [5] and 8.72% more than that of [13].

The packet delivery ratio is measured for fixed 250 nodes in the network and by varying the attack rate from 5% to 20% and the result is shown in Table III and plotted in Fig. 5.

TABLE I. Parameters of Simulation

Parameters	Values
Number of Nodes	50 to 250
Transmission range(m)	100
Simulation area(m^2)	1000*1000
Node propagation	Random
Span of Simulation (minutes)	30
Queue Size of Interface	50
Medium Access Control	802.11
Percentage of attackers	10% of total nodes

TABLE II. Packet Delivery Ratio

No of Nodes	Proposed	[5]	[13]
50	88.4	82.15	81.5
100	90.2	83.34	82.17
150	91.5	84.56	83.11
200	92.7	85.12	84.32
250	93.6	86.31	85.5

Fig. 4. Packet Delivery Ratio.

TABLE III. Packet Delivery Ratio based on Attack Percentage

Percentage of attacker	Proposed	[5]	[13]
5	96.4	88.15	87.5
10	93.6	86.31	85.5
15	91.5	84.56	83.11
20	90.7	83.12	82.32

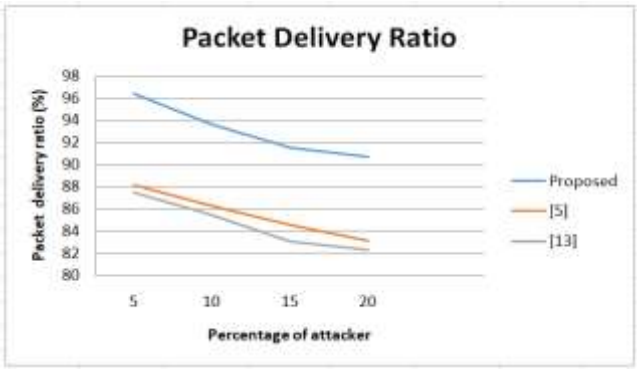


Fig. 5. Packet Delivery Ratio based on Attack Percentage.

In WSN, the packet delivery ratio decreases as the attack rate increases. But the packet delivery ratio is still higher in the proposed solution. It is 8.07% higher compared to [5] and 9.08% higher compared to [13].

The accuracy of attack detection is measured by varying the number of nodes with 10% of nodes as attacker and the result is presented in Table IV.

The attack detection ratio in the proposed solution is 7.72% higher compared to [5] and 8.82% higher compared to [13]. The reason for increased attack detection ratio is due to localization of detection to zone level in the proposed solution.

False positives are very common in any detection technique. Certain drops due to network conditions could be wrongly misinterpreted as message dropping attack. False positives are measured by varying the number of nodes with 10% of nodes as attacker and the result is given in Table V and Fig. 6.

The false positives in the proposed work is 28% lower compared to [5] and 18.3% lower compared to [13]. The reason for reduced false positives it due to better watch dog mechanism with localized monitoring and sequential probability test in the proposed solution.

TABLE IV. ACCURACY OF ATTACK DETECTION

No of Nodes	Proposed	[5]	[13]
50	90.4	83.25	82.5
100	91.4	84.44	83.17
150	92.5	85.51	84.11
200	93.6	86.15	85.32
250	94.5	87.33	86.5

TABLE V. FALSE POSITIVES

No of Nodes	Proposed	[5]	[13]
50	10.4	13.65	12.5
100	11.5	14.54	13.17
150	12.2	15.61	14.71
200	13.3	16.75	15.82
250	13.8	17.83	16.2

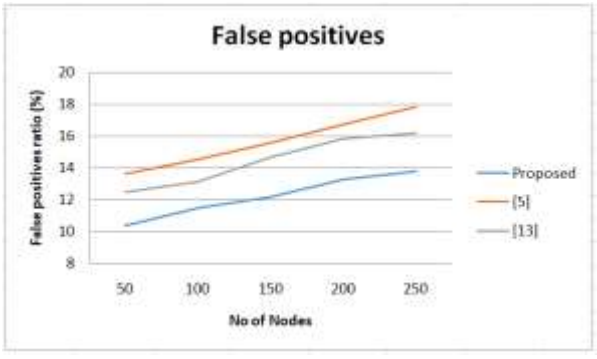


Fig. 6. False Positives.

Time for detection of attack is measured for a fixed node of 250 by varying the percentage of attacks and the result is given in Table VI and Fig. 7.

The time for detection of attack is almost flat with only a slight increase in the time compared to [5] and [13]. This is because of parallelization is detection at zone level.

The network overhead is calculated for a fixed node of 250 by varying the percentage of attacks and the result is given Table VII and Fig. 8.

TABLE VI. ATTACK DETECTION TIME

Percentage of attacker	Proposed	[5]	[13]
5	11	13	12
10	12	15.54	14.17
15	12.5	17.61	16.71
20	12	20.75	18.82
25	12.8	22.83	21.2

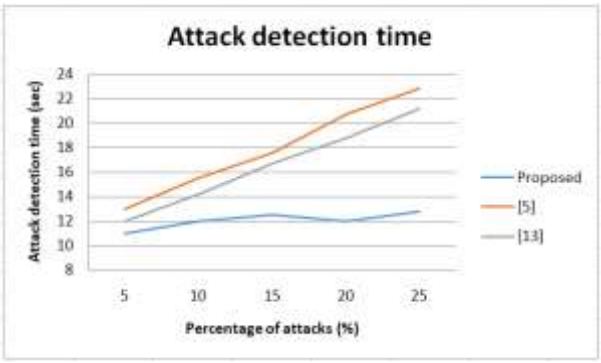


Fig. 7. Attack Detection Time.

TABLE VII. NETWORK OVERHEAD

Percentage of attacker	Proposed	[5]	[13]
5	8	13	12
10	11	15	14
15	15	18	17.2
20	19	22	21
25	22.5	24	23.2

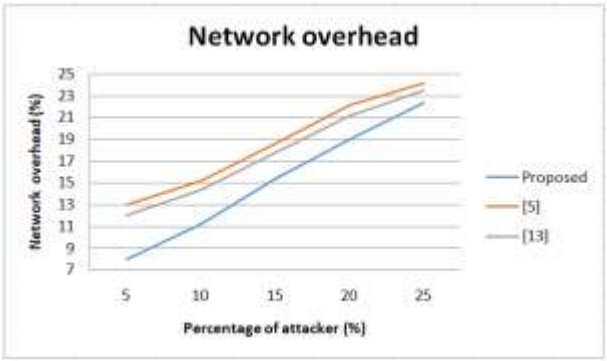


Fig. 8. Network Overhead.

The network overhead in the proposed work is 22.52% lower compared [5] and 17% lower compared to [13]. The proposed solution has lower overhead because of the distributed nature in the proposed solution and attack influence is localized.

V. Conclusion and Future Work

In this work, a secure intruder information sharing in wireless sensor network for attack resilient routing is proposed. The proposed solution is able to detect message drop, message tampering and message flooding attacks with higher accuracy when compared to existing solutions. The malicious node is identified and isolated in the zone. Also due to attack detection localization with zones, the network overhead and time to detect attack is comparatively lower in the proposed solution. The information about the attacker is shared in a secure manner using HECC and there is higher reliability for attacker information sharing in the network.

Due to unattended node deployment batteries may have limited power which requires additional resources to recharge.

As part of the future work, the proposed work can be extended to increase communication range in secured manner with Realtime scenario by considering collision avoidance and energy constraints.

References

[1] Venkateswara Rao M and Srinivas Malladi, “Secure Energy Efficient Attack Resilient Routing Technique for Zone based Wireless Sensor Network” International Journal of Advanced Computer Science and Applications (IJACSA),11(12),2020.http://dx.doi.org/ 10.14569 / IJACSA.2020.0111267.

[2] Stefanos A. Nikolidakis, Dimitrios D. Vergados, Christos Douligeris Algorithms, Energy Efficient Routing in Wireless Sensor Networks Through Balanced Clustering, 2013.

[3] J. Jiang, Y. Liu and B. Dezfouli. (2018): A Root-based Defense Mechanism Against RPL Blackhole Attacks in Internet of Things Networks, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018,pp.1194-1199,doi:10.23919/APSIPA. 2018.8659504.

[4] M.Rajesh Babu,S. Moses Dian, Siva Chelladurai, Mathiyalagan Palaniappan(2015) : Proactive Alleviation Procedure to Handle Black Hole Attack and Its Version, The ScientificWorld Journal, vol.2015,ArticleID 715820,11 pages, 2015. https://doi.org/10.1155/2015/715820.

[5] Zhang, Qiong & Zhang, Wenzheng. (2019). Accurate detection of selective forwarding attack in wireless sensor networks. International Journal of Distributed Sensor Networks. 15. 155014771882400. 10.1177/1550147718824008.

[6] Rutvij H. Jhaveri and Narendra M. Patel,"A sequence number based bait detection scheme to thwart gray hole attack in mobile ad hoc networks”, Wireless Network, Springer, 2015, Vol.21, Issue 8, pp.2781-2798.

[7] Y M. Khamayseh,ShadiA,Aljawarneh, and Alaa Ebrahim Asaad, “Ensuring Survivability against Black Hole Attacks in MANETS for Preserving Energy Efficiency”, Sustainable Computing: Informatics and Systems. Vol.18, pp.90-100, suscom.2017.07.001 http://dx.doi.org/ 10.1016/j.

[8] S. S. Albouq and E. M. Fredericks.(2017): Lightweight Detection and Isolation of Black Hole Attacks in Connected Vehicles, IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW), Atlanta, GA, 2017, pp. 97-104, doi: 10.1109/ICDCSW.2017.23.

[9] T. Poongodi and M.Karthikeyan ,"Localized Secure Routing Architecture Against Cooperative Black Hole Attack in Mobile Ad Hoc Networks", Wireless Personal Com.,Springer,2016.Vol.90, Issue 2, pp.1039-1050.

[10] R. Baiad, H. Otrok, S. Muhaidat and J. Bentahar. (2014) :Cooperative cross layer detection for blackhole attack in VANET-OLSR, International Wireless Communications and Mobile Computing Conference (IWCMC), Nicosia.

[11] Hamamreh, Rushdi. (2018). Protocol for Multiple Black Hole Attack Avoidance in Mobile Ad Hoc Networks. 10.5772/intechopen.73310.

[12] Ismail,Asima & Amin, Rashid. (2019). Malicious Cluster Head Detection Mechanism in Wireless Sensor Networks. Wireless Personal Communications. 108. 10.1007/s11277-019-06512-w.

[13] Terence,Sebastian& Purushothaman, Geethanjali. (2019). A novel technique to detect malicious packet dropping attacks in wireless sensor networks. Journal of Information Processing Systems. 15. 203-216. 10.3745/JIPS.03.0110.

[14] Wazid, Mohammad & Katal, Avita & Sachan, Roshan & Goudar, R.H. & Singh, Dharam. (2013). Detection and prevention mechanism for Blackhole attack in Wireless Sensor Network.576-581. 10.1109/iccsp.2013.6577120.

[15] Shafiei, Hosein &Khonsari, Ahmad & Derakhshi, Hazhir & Mousavi,P..(2014). Detection and mitigation of sinkhole attacks in wireless sensor networks.Journal of Computer and System Sciences.80.644–653.10.1016/j.jcss.2013.06.016.

[16] Wazid, Mohammad & Das,Ashok Kumar. (2016). A Secure Group-Based Blackhole Node Detection Scheme for Hierarchical Wireless Sensor Networks. Wireless Personal Communications.94.10.1007/s11277-016-3676-z.

[17] Manjunath, B. E., and P. V. Rao. "Trends of Recent Secure Communication System and its Effectiveness in Wireless Sensor Network." Journal of Innovation in Electronics and Communication Engineering 6.2 (2016): 46-52.

[18] Dr. P.V.Rao, Manjunath B E, “Unique Analytical Modelling of Secure Communication in Wireless Sensor Network to Resist Maximum Threats”, International Journal of Advanced Computer Science and Applications, (IJACSA) Vol. 10, No. 2, pp 421-427, 2019.

[19] Chowdary, Krishna, and K. V. V. Satyanarayana. "MALICIOUS NODE DETECTION AND RECONSTRUCTION OF NETWORK IN SENSOR ACTOR NETWORK." Journal of Theoretical & Applied Information Technology 95.3 (2017).

[20] Vamshi krishna, H., & Swain, G. “Identification and avoidance of malicious nodes by using certificate revocation method.” International Journal of Engineering and Technology (UAE), 7(4.7 Special Issue 7) (2018).

[21] Prasad,A.Y.,and Balakrishna Rayanki."A generic algorithmic protocol approaches to improve network life time and energy efficient using combined genetic algorithm with simulated annealing in MANET." International Journal of Intelligent Unmanned Systems (2019). Vol. 8 No. 1, pp. 23-42.

[22] Prasad,A.Y.and R.Balakrishna. "Implementation of optimal solution for network lifetime and energy consumption metrics using improved energy efficient LEACH protocol in MANET." Telkomnika Vol. 17 No.4 (2019): 1758-1766.

Mobile Technologies' Utilization and Competency among College Students

Mokhtar Hood Bindhorob[1]
Department of Education Technology
College of Educational Studies
University Putra Malaysia

Khaled Salmen Aljaaidi[2]
Department of Accounting
College of Business Administration
Prince Sattam bin Abdulaziz University

***Abstract*—The focus of this study is to (1) determine the level of mobile technologies' utilization and competency and (2) report separately the level of basic operation, communication and collaboration, information Seeking, digital citizenship and creativity and innovation skills in mobile technologies among first year undergraduate students in College of Computer Science and Information Technology at Hadhramout University. The sample size consists of 148 freshmen students. Using the descriptive statistical analysis, the results of this study reveal that undergraduates in College of Computer Science and Information Technology have highly utilized mobile technologies. It was also figured out that they were extremely capable to use these devices. Moreover, it was revealed that undergraduates' competency and utilization levels were so high in communication purpose due to certain social and educational reasons. Due to the results of the study, wider and greater implication of this method in College of Computer Science and Information Technology and other colleges at Hadhramout University is recommended along with the activation of the university apps.**

Keywords*—Mobile technologies; utilization; competency; college students

I. Background of the Study

World Bank Institute reported that the number of smart phone users in 2010 had increased to 5 billion as compared to only 700 million users in early year of 2000, which indicates that the increase of users in percentages would be more than 70% of the world population in future. With that figure, it would absolutely be doubted whether underdeveloped country like Yemen is on the race with other developed countries or not. A report from [15] declared that the number of mobile customers in Yemen increased in December 2015 to reach to 16.88 million users. In this same report, it is declared that the basic mobile phone facilities were steadily becoming familiar with mobile networks reaching about 90 percent of the population. Moreover, it is discovered that the number of mobile phone subscriptions in Yemen had notably raised to five-fold from 3 million in 2006 to 16 million in 2016, resulting to 56.9-percentage penetration for a population of about 28 million. Furthermore, the report detailed that the claim for internet-enabled 3G facilities was increasing as Yemenis started switching from using basic mobile phones to smartphones and computers such as laptops and tablets [1].

This growth is expected to have an effect of further learning evolution, particularly in higher education institutions (HEI). These institutions will be going through quick changes, especially during the next 10 years due to this rapid growth [15]. Moreover, it is obvious that ubiquitous information systems are altering the creation and dissemination of information in new ways, producing opportunities in all aspects of society [16] [11]. In their study; [13] consider mobile learning as a major evolution in HEI learning sector where students and professors are more enlightened in technology due to its existence. On the other hand, the author in [17], revealed that HEI has extremely modified the reason of the continuous development of mobile computing devices and internet resources. The author in [14] stated that HEI students might be ready to accept m learning sooner than K-12 students do because more college students have their own mobile devices. The author in [15], concluded in their survey, which is conducted in Republic of Yemen, that students in HEI have positive attitude towards the use of e-learning and m-learning. They also discovered that they are familiar with the use of smartphone facilities in daily activities.

However, the availability of mobile devices does not necessary ensure high utilization of the devices among students particularly in education field. We must first assess the students' readiness for mobile learning [2] [5] [6] with regard to these issues. The author in [15] discovered that despite the familiarity of e learning and m learning among undergraduate students in Yemen, HEI students do not use smartphones for learning actively. Therefore, the aims of this study are to (1) determine the level of mobile technologies' utilization among first year undergraduate students, (2) identify the level of mobile technologies' competency among first year undergraduate students, and (3) report separately the level of basic operation, communication and collaboration, information Seeking, digital citizenship and creativity and innovation skills in mobile technologies competency.

The importance of this study stems out of the fact that mobile learning through mobile technologies will be the centre focus of this study because Yemen is one of the countries that experiencing a scarcity of mobile technology resources, where the utilization of mobile technologies is known to be quite limited. Moreover, this study also attempts to reveal the level of mobile technologies utilisation among the students. Hence, it is hoped that the results of this study will later bring on benefits to the higher authority of the university, to improve the existing learning conditions. Another worth noted point is that this study would later on be a good reference to other future studies within the same context in Yemen or in any Arabic country. With very little searched topics in the field of

integrated technologies especially in mobile technologies adoption, it is vital and essential to tackle this topic.

II. Research Methodology

A. Population and Samples

In a country like Yemen where integration of technology is in an infant stage, it becomes very distinguished, remarkable and of a great interest to have a fully equipped faculty. In college of computer science and information technology, technologies like mobile devices are extensively applied. Therefore, this faculty have been chosen for this study to investigate the influence of competency of mobile technologies among the undergraduate students on their frequent utilization of these devices. Another possible reason is that majority numbers of staff in this faculty possesses an outstanding knowledge and are among specialists in the field of computer science and information technologies. All of the staff in this faculty is continuously trying to innovate their teaching method in line with the use of current and available technologies.

Besides, using mobile devices like laptops, tablets and mobiles inside classes are compulsory for all students in this faculty to facilitate the learning process. Majority of the given task need to be done outside classes, purposely to let the students experience the mobile learning approaches. More precisely, all first year students in Computer Science and Information Technology Departments in this faculty are selected to participate in the study due to some of characteristics found in them. First, these students are newly introduced to new purposes of using these devices as stated before and uses of mobile technologies especially learning, studying and research purposes. Moreover, they are studying subjects with a relation to the topic of this study like; communication skills, learning skills, thinking skills and searching skills in the second semester of their first year. They are even studying about theory of computation. Hence, they possess at least the basis knowledge of M-Learning in general. Moreover, the choice of time which is the second semester when it is applicable to check whether being in this department studying these subjects and knowing the new purposes changed the way they utilize these devices.

In addition to that, the researcher wants to check whether the availability of these technologies and the allowance of using them will affect the way these undergraduates perceive the technologies and use them. Due to the above-mentioned reasons, all of the first year undergraduates were selected as the study sample, which means a census population was chosen. This also means that no sampling was done for this study. Moreover, the selection of the sample of this study is supported by study conducted by (Dawson, 2007; Saunders et al., 2007; Zikmund et al., 2010) who states that census is the most suitable for this research since participants are so rare and distinguished. They are of great interest to the researcher since they have owned previously stated features.

B. Instrument

Due to the characteristics of questionnaire survey which offers cheap and time saving, hence the instrument is distributed with online and offline method. Moreover, the results of the survey questionnaire can be generalized [4]. The instrument used in this study is divided into three sections: Section 1: The demographic information, Section 2: the utilization of mobile technologies, and Section 3: the competency factor. Table I presents the summary of the survey instruments sections.

1) Demographic information: The items are developed by the researcher to get respondent's personal information particularly on the history of these technologies. There are four items or questions in this section. All of them are close-ended questions. All these items are asking about personal information like, departments (CS or IT), gender, and age. Etc.

2) Utilization of mobile technologies: This part of the instrument is adopted from [12] which quoted from [7]. This section contains 40 items which divided into six sub sections of: i) basic information uses, ii) communication uses, iii) collaboration uses, iv) information seeking uses, v) digital citizenship uses, and vi) creativity and innovation uses. The aim of this section is to explore the undergraduate students' level of utilization of mobile technologies. In other words, it will reveal how frequently they utilized the mobile application in related to M-Learning. Each item in every subsection is measured through five Likert scale which represented by (1) to (5), (1) never, rarely (Once or twice), (2) sometimes (2 – 3 times a week), (3) often 4 – 6 times a week and (5) almost every day. Table II shows the six subsections in the utilization of mobile technologies section with the number of items for each section.

TABLE I. Sections in the Questionnaire

Section	Aspect measure	No. of items
Section 1	Demographic information	5
Section 2	Utilization of mobile technologies	40
Section 3	Competency factor	40

TABLE II. Subsections or Dimensions with Items of Mobile Technologies Utilization

Name of subsection	No. of items
Basic operation uses	7
Communication uses	9
Collaboration uses	3
Information seeking uses	7
Digital citizenship uses	9
Creativity and innovation uses	5

C. Data Collection

There are few challenges that need to be dealt with during the data collection phase. Despite the easiness of distributing questionnaires online, it is notably hard to administer it. Therefore, the researcher prefers to distribute the questionnaires in a collective way. Having a good relationship with lecturers was another reason why the researcher prefers to perform it in collective questionnaire as compared to the online version. Majority of the academicians in the faculty are helpful and collaborative enough to lend a hand in the survey distribution phase. The undergraduate student's weak experiences of using email and online questionnaire was another reason for the researcher to choose the collective method.

The data collection step was carried out from August to September 2018. A total number of 208 undergraduate students had participated in this study. Thirty of them were chosen for the pilot study. The rest of them were used for the actual study, which were 178. Out of the number of questionnaires distributed, 162 were returned to the researcher. After checking the questionnaires responses, 148 of the 162 of the returned were valid, completed and used for the analysis, which is a good feedback. The number of questionnaires that are valid completed and returned are more than 60% of the total questionnaires distributed. This is considered a good number for the data analysis and report [3]. Table III shows the response rates and percentages on survey questionnaire distributed.

When the questionnaires were collected, sorted and checked, the data was entered to IBM SPSS 23 statistical package for analysis.

TABLE III. RESPONSE RATES AND PERCENTAGES ON QUESTIONNAIRE

Description	Distributed in actual study	Returned	Valid
No. of questionnaire	178	162	148
Percentage	100%	91%	83%

III. RESULTS AND DISCUSSIONS

A. Demographic Information

This first section of the survey questionnaire is on respondent demographic information. The demographic information involved department, sex, age, mobile technologies own, frequency of mobile technologies usage and hours spend on the device involved. A total number of 148 respondents involved in this survey. Information is presented in Table IV.

Table IV indicates from 148 respondents, majority of them are from IT department (50.7), 49.3% are from CS department. Meanwhile, majority of the respondents are male with 62.2 % as compared to female with 37.8%. The number of male respondents is more than females because in most of Arab countries women' right is still not fulfilled specially in education. Therefore, the number of females in HE institutions still smaller if compared to males. Majority of them with 80.4 % age from 20-24 years old, followed by 15-19 years old (13.5%), and 25-30 with 6.1 %. Most of the respondent age from 20-24 years old is due to the education systems, which force students to start studying at the age of seven and then twelve years for primary and secondary. Following that, students are forced to have one-year vacation. When calculating, it is discovered that early twenties is the age for majority of students in the first year unless there are special cases. Regarding the mobile technologies ownership, majority of them own a laptop (81.8%), while the rest does not own a laptop (18.2%).

Meanwhile, all of the respondents own a mobile or smart phone. A total number of 85.8% of the respondent own a tablet, while only 14.2% among them does not own a tablet. Based on the output on mobile technologies usage, most of the respondents are using the laptop for 2-5 years (33.8%), followed by months -1 years (23.6%), never (23.0%) and over 5 years (19.6%). Majority of them are using the smartphone more than 5 years (49.3%), followed by 2-5 years' usage (38.5%), and the least is month-1 year (12.2%). Surprisingly, majority number of the respondent never uses a tablet. This is due to the fragility of tablets and quickly damaged. While another 8.8% had use them over 5 years, followed by 2-5 years (7.7%) and the least is months-1 year (2.7%).

The results on daily usage of each devices reveals that the respondents commonly spend approximately 1-5 hours (64.9%) on the laptop usage, whereas only few of them spend more than 10 hours on the laptop usage. The smart phone is commonly utilized daily with approximately 1-5 hours (45.3%). Lastly, only 9.5% of the respondents spend they daily time on tablet. This is because majority number of the respondents does not own a tablet.

B. Overall Utilization of Mobile Technologies among Undergraduates

This section is design to identify the level of mobile technologies utilization based on user's frequency of using mobile technologies. In this section the question "What is the level of utilization among first year undergraduate students when using mobile technologies?" will be answered in this part of the results. To answer this question, the respondents were asked to indicate their level of frequency on each purpose of mobile technologies utilization which was measured by the five Likert scale; (1) never, (2) Rarely (once or twice), (3) Sometimes (2 – 3 times a week), (4) Often (4 – 6 times a week) and to the most frequent use (5) almost every day. The results presented involved 148 respondents respectively. The results of all dimensions in the dependent variables will be stated like in the Table IV.

As mentioned earlier, one of the objectives of this study is to study on the level of mobile technologies utilization among CS & IT undergraduates' students in general. Therefore, based on Table V, the output had confirmed that out of six factors of mobile technologies utilization purposes, majority of the undergraduates' students frequently used the mobile technologies by means of communication (M =4.6, SD = 2.62). Following that, basic operation purposes (M =4.0, SD = 4.02) and information seeking (M =4.0, SD = 3.50) are equally and highly utilized by HE students. Creativity and innovation with (M =3.56, SD = 2.46) is in the third rank.

Lastly, Digital citizenship with (M =3.0, SD = 4.88) and collaboration purpose (M =3.0, SD = 2.39) are as noticed the least equally utilized among the students.

TABLE IV. RESPONDENT'S DEMOGRAPHIC

Variables	Category	Frequency	Percentage
Department	CS IT	73 75	49.3 50.7
Sex	Male Female	92 56	62.2 37.8
Age	15 – 19 20 – 24 25 – 30	20 119 9	13.5 80.4 6.1
Own laptop	Yes No	121 21	81.8 18.2
Own mobile	Yes No	148 -	100 -
Own tablet	Yes No	21 127	14.2 85.8
Period of using laptop	Never months - 1 year 2 years - 5 years over 5 years	34 35 50 29	23.0 23.6 33.8 19.6
Period of using smartphone	Never months - 1 year 2 years - 5 years over 5 years	- 18 57 73	- 12.2 38.5 49.3
Period of using tablet	Never months - 1 year 2 years - 5 years over 5 years	124 4 7 13	83.8 2.7 7.7 8.8
Period of using laptop (daily)	Never 1 hour - 5 hours 6 hours - 10 hours More than 10 hours	32 96 12 8	21.6 64.9 8.1 5.4
Period of using smartphone (daily)	Never 1 hour - 5 hours 6 hours - 10 hours More than 10 hours	- 67 44 37	- 45.3 29.7 25.0
Period of using tablet (daily)	Never 1 hour - 5 hours 6 hours - 10 hours More than 10 hours	131 14 3 -	88.5 9.5 2.0 -

TABLE V. THE LEVEL OF MOBILE TECHNOLOGIES UTILISATION AMONG CS AND IT UNDERGRADUATE STUDENTS

Items	N	Mean	SD	Rank
Communication	148	4.6	2.62	1
Basic operation Purposes	148	4.0	4.02	2
Information Seeking	148	4.0	3.50	2
Creativity and Innovation	148	3.56	2.46	3
Digital Citizenship	148	3.0	4.88	4
Collaboration	148	3.0	2.39	4

C. Overall Competency of Mobile Technologies among Undergraduates

This section is designed to identify the level of mobile technologies competency based on user's skilfulness of using mobile technologies. In this section, the question "What is the level of competency among first year undergraduate students when using mobile technologies?" will be answered in this part of the results. To answer this question, the respondents were asked to indicate their level of skilfulness on each purpose when using these devices which was measured by the five Likert scale: (1) No skills in this area, (2) Limited skills in this area, (3) enough Skills: Need refinements, (4) Skilful (5) Very Skilful. The results presented involved 148 respondents respectively.

The results in Table VI clearly displays that the most prominent competency level among the students is as expected basic operation tools (M =3.9, SD = 3.49), followed by information seeking tools (M =3.8, SD = 3.10). Thirdly, HE students are competent in communication purposes (M =3.7, SD = 3.78). Meanwhile, creativity and innovation tools is unexpectedly the next mastered skill among the students (M =3.4, SD = 2.52). The HE students are least skilled on digital citizenship (M =3.0, SD = 4.40) and collaboration (M =2.9, SD = 2.07).

Table VII demonstrates the results on the overall results of utilization and competency of mobile technologies in the learning field conducted among the undergraduates' students. Based on the above table, it can be clearly seen that the results of overall utilization of mobile learning is (M = 3.78, SD =12.1). Meanwhile, the overall competency level is (M= 3.53, SD= 13.5). Hence, the results had indicated that CS & IT undergraduates in Hadhramout University are highly utilizing the mobile technologies in their hands. Moreover, it is indicating that these students are so competent in using these devices for the six purposes.

TABLE VI. RESULTS ON OVERALL COMPETENCY LEVEL OF MOBILE TECHNOLOGIES

Items	N	Mean	SD	Rank
Basic operation Tools	148	3.9	3.49	1
Information Seeking Tools	148	3.8	3.10	2
Communication Purposes	148	3.7	3.78	3
Creativity and Innovation Tools	148	3.4	2.52	4
Digital Citizenship Tools	148	3.0	4.40	5
Collaboration Purposes	148	2.9	2.07	6

TABLE VII. OVERALL; MEAN SCORE AND STANDARD DEVIATION FOR MOBILE LEARNING UTILIZATION AND COMPETENCY LEVEL

Dimension	N	Mean
Overall utilization	148	3.78
Overall competency	148	3.53

D. Individual Level of Utilization and Competency for all Dimensions among Undergraduates

Now a detailed description about which of the statements in the dimensions are higher will be displayed in the following tables. In this section the question "What is the level of basic operation, communication and collaboration, information Seeking, digital citizenship and creativity and innovation skills in mobile technologies competency and utilization separately?" will be answered. First, detailed results for all dimensions in utilization of mobile technologies will be provided. Second, the results will detail data about all dimensions in competency of mobile technologies.

1) All dimensions in utilization of mobile technologies: Table 8 above demonstrates the results on mobile technologies for basic operation purposes. According to previous literature review, seven significant dimensions identified were included under mobile technologies for basic operation purposes. These dimensions are: i) sharing files and documents, ii) setting time and place for an event, iii) capturing pictures, iv) recording videos, v) installing applications, vi) organizing files into folders and vii) opening several programs simultaneously. Based on the compare of mean results, it can be clearly seen that the most highly utilize mobile technology under this factor which rank as the first is operating several program simultaneously (M = 4.34, SD = .676), followed by organizing files into folders (M = 4.09, SD = .755). Meanwhile, the third highest utilization ranked is recording videos (M = 4.07, SD = .809), followed by (setting time and place for an event (M = 4.05, SD = .875), installing application (M = 4.01, SD = .861), capturing pictures (M = 3.90, SD = .855) and the least utilized among the undergraduate's students of this faculty is sharing files and documents (M = 3.08, SD = .849).

In this subsection, the respondent was asked to indicate their level of frequency of communication with three groups of people that are friends, lecturer, and family members with similar scale of 1 to 5 as previous aforementioned section. Each table below demonstrates the results generated on the level of utilization of mobile technologies using SPSS.

Table IX displays the results on communication purposes using mobile technology among the respondent with their friends. Three significant communication medium used as dimension in this subsection are: i) voice, ii) video and iii) email. Based on the above table, the results proved that the most highly utilize mobile technologies under this factor is using email (M = 4.61, SD = .504), followed by video (M = 4.57, SD = .510) and lastly voice (M = 4.54, SD = .539). Voice is the most preferred communication medium, because it is more expressive especially when they are very busy. Meanwhile, email is the least preferred communication medium with their friends, because most of what they are sharing is talks and chats. Moreover, they do have the same way of thinking that emails are meant for official matters only.

TABLE VIII. MOBILE TECHNOLOGIES FOR BASIC OPERATION PURPOSES RESULTS

Item	**N**	**Min**	**Max**	**Mean**	**SD**	**Rank**
Opening several programs simultaneously	148	3	5	4.34	.676	1
Organizing files into folders	148	3	5	4.09	.755	2
Recording videos	148	2	5	4.07	.809	3
Setting time and place for an event	148	2	5	4.05	.875	4
Installing applications	148	1	5	4.01	.861	5
Capturing pictures	148	2	5	3.90	.855	6
Sharing files and documents	148	1	5	3.86	.849	7

TABLE IX. MOBILE TECHNOLOGIES FOR COMMUNICATION PURPOSES RESULTS (FRIENDS)

Item	N	Min	Max	Mean	SD	Rank
UTICF (Email)	148	3	5	4.61	.504	1
UTIF(Video)	148	3	5	4.57	.510	2
UTIF (Voice)	148	3	5	4.54	.539	3

TABLE X. MOBILE TECHNOLOGIES FOR COMMUNICATION PURPOSES RESULTS (LECTURER)

Item	N	Min	Max	Mean	SD	Rank
UTICL (video)	148	3	5	4.65	.507	1
UTICL (email)	148	3	5	4.62	.514	2
UTIL (voice)	148	3	5	4.59	.521	3

Table X indicates the results on communication purposes using mobile technology among the respondent with their lecturers. Three significant communication medium used as dimension in this subsection are: i) voice, ii) video and iii) email. Based on the above table, the results demonstrate that the most highly used communication medium that students used to communicate with the lecturers is video (M = 4.65, SD = .507), followed by email (M = 4.62, SD = .514) and lastly voice (M = 4.59, SD = .521). As compared to previous results of which they prefer to communicate with their friends using email, majority of the students prefers to communicate using video with their lecturer. This is due to the mode of teaching and studying. Some of the teachers attempt to let the students experience different modes of delivering information. Therefore, they were videotaping their lectures. Furthermore, students sometimes are asked to video tape themselves when doing some of their assignments. Moreover, voice is the least preferred communication medium is because it not much required by their lecturer when they are asking them to do an assignment. Note that the communication is not face-to-face type rather than it is recorded videos.

Table XI shows the results on communication purposes using mobile technology among the respondent with their family. Based on the above table, the results indicate that the most highly used communication medium that undergraduates used to communicate with their family members is using voice (M = 4.60, SD = .518), followed by video (M = 4.59, SD = .520) and lastly email (M = 4.59, SD = .521). The results clearly stated that respondents prefer to communicate with their family using voice as compared to video and email. Voice is the most preferable communication medium because videos are difficult to be used if compared. The videos would have been the most preferred unless the bandwidth is very terrible in the countryside (most of the students are either form KSA or countryside). They feel that text is not fulfilling what they want to say and videos are hard to use. Then, it is only through the voice that they, with little difficulty, can send voice mails rather than calling.

Table XII demonstrates the results on the level of utilization of mobile technologies for collaboration purposes. The results indicate that the most frequently used mobile technologies for collaboration is delivering a presentation via a videoconference (M =3.08, SD = .973), followed by attending a webinar (M =2.97, SD = .979), and the least use is sharing ideas via other people's blogs, social forum, social networking sites' walls etc. (M =2.82, SD = .822). The students frequently use to deliver a presentation via a videoconference as compared to two other factors is mainly due to the well –equipped labs. Due to the university, powerful internet bandwidth, which is wireless, lecturers, experiences the video conferences with their friends from other countries. It is one of the new styles of teaching in this outstanding college. That happens only inside the college.

Table XIII displays the output on the level of utilization of mobile technologies for information seeking purposes. The results prove that the most frequently used mobile technologies under the information seeking categories factor is to get latest information from subscribing portal/ website (M =4.14, SD = .753). The next result is the one of getting latest information by signing up for newsletter of portals and websites (M =4.07, SD = .739). After that is finding latest information on subject / topic by browsing with (M =4.03, SD = .884). Then updating list of portals and websites in bookmark (M =3.92, SD = .796) is following. Bookmarking portals and website related to the field (M =3.89, SD = .905) is the next. After that, browsing to find references (M =3.84, SD = .825) is the next. The least utilized information using mobile technologies among the students is downloading learning materials from portals and websites (M =3.80, SD = .873). The students frequently utilized the mobile technologies to get latest information from subscribing portal/ website mainly because there is always new information in computer science and information technology field offered in portals and websites guaranteed by the lecturers. Moreover, lecturers in this college keep always asking them to collect info from the authorized portals and websites to keep them up to date.

Table XIV presents the result on the level of utilization of mobile technologies for digital citizenship purposes. The results have proven that the most frequently used mobile technologies under this factor is to using other people' works lawfully (M =3.49, SD = .951). It is followed by learning and prevent on cyber bullying (M =3.22, SD = 1.07) and expressing myself through digital media (M =3.16, SD = .931). After that, it is the item "creating strong passwords to protect my private information (M =3.08, SD = .944)" and "protecting my digital works through the copyright (M =2.96, SD = .790)". Next, is the items "finding out required information (M =2.86, SD = .901) and "evaluating information (M =2.84, SD = .886)". The least frequently used are "identifying valuable information (M =2.81, SD = .884)" and "using information effectively (M =2.71, SD = .859)". Using other people's work lawfully is the most preferable is due to the continuous warnings given to them by their lecturer. When the researcher asked the lecturers about this issue, their answer was "we give them stories happened in the countries we studied in to draw the awareness to this issue".

TABLE XI. MOBILE TECHNOLOGIES FOR COMMUNICATION PURPOSES RESULTS (FAMILY)

Item	N	Min	Max	Mean	SD	Rank
UTICFA(voice)	148	3	5	4.60	.518	1
UTICFA(video)	148	3	5	4.59	.520	2
UTICFA(email)	148	3	5	4.59	.521	3

TABLE XII. MOBILE TECHNOLOGIES FOR COLLABORATION PURPOSES

Item	N	Min	Max	Mean	SD	Rank
Delivering a presentation via a video Conference	148	1	5	3.08	.973	1
Attending a webinar	148	1	5	2.97	.979	2
Sharing ideas via other people's blogs, social forum, Social Networking Sites' walls etc	148	1	5	2.82	.822	3

TABLE XIII. MOBILE TECHNOLOGIES FOR INFORMATION SEEKING PURPOSES

Item	N	Min	Max	Mean	SD	Rank
Get latest information from subscribing portal/ website	148	2	5	4.14	.753	1
Get latest information by signing up for newsletter of portals and websites	148	2	5	4.07	.739	2
Finding latest information on subject / topic by browsing.	148	2	5	4.03	.884	3
Updating list of portals and websites in bookmark	148	1	5	3.92	.796	4
Bookmarking portals and website related to the field	148	1	5	3.89	.905	5
Browsing to find references	148	1	5	3.84	.825	6
Downloading learning materials from portals and websites	148	1	5	3.80	.873	7

TABLE XIV. MOBILE TECHNOLOGIES FOR DIGITAL CITIZENSHIP PURPOSES

Items	N	Min	Max	Mean	SD	Rank
Using other people' works lawfully	148	2	5	3.49	.951	1
Learning what to do if involved in cyberbullying & how to overcome	148	1	5	3.22	1.07	2
Expressing myself through digital media	148	1	5	3.16	.931	3
Creating strong passwords to protect my private information	148	1	5	3.08	.944	4
Protecting my digital works through the copyright	148	1	5	2.96	.790	5
Finding out required information	148	1	5	2.86	.901	6
Evaluating information	148	1	5	2.84	.886	7
Identifying valuable information	148	1	5	2.81	.844	8
Using information effectively	148	1	5	2.71	.859	9

TABLE XV. MOBILE TECHNOLOGIES FOR CREATIVITY AND INNOVATION PURPOSES

Items	N	Min	Max	Mean	SD	Rank
Uploading work	148	1	5	3.61	.779	1
Creating an innovation/product	148	2	5	3.59	.832	2
Producing work	148	2	5	3.57	.661	3
Performing advanced searches	148	2	5	3.51	.705	4
constructing an original work	148	2	5	3.50	.655	5

Referring to Table XV, the results on mobile technologies for creativity and innovation purposes shows that uploading work is the most frequently perform under this factor (M =3.61, SD = .779). Meanwhile, the second highest preferred is by creating an innovation/product (M =3.59, SD = .832), followed by producing work (M =3.57, SD = .661), performing advance searches (M =3.51, SD =705), and lastly the students choose to utilized the mobile technologies by constructing an original work (M =3.50, SD = .655). The students highly utilized the mobile technologies by uploading their works due to the assignment that they have to do and then upload to their face book group to discuss it with their peers.

Finally, it could be concluded that_communication is the most highly utilized by the students is due to certain reasons. These students are far from their families; therefore, they need to contact their families from time to time. More importantly, these students are given assignments to answer in a due time. For this purpose, students may contact their lecturers to have a clear picture about the assignment or they may ask their classmates instead. This notion is supported by the study conducted by Murphy (2011). In this study, it is revealed that students are very energetically using technologies available for communication purposes to facilitate their learning process through interactions.

2) All dimensions in competency of mobile technologies: Table XVI shows the results of competency of mobile technologies as basic operation tools. The above table indicates that the most highly rated skill is opening several programs simultaneously (M = 4.20, SD = .670), followed by organizing files into folders (M = 3.97, SD = .713), recording video (M=3.96, SD =.737), setting time and place for an event (M=3.95, SD =.772), sharing files and documents (M=3.78, SD =.787), installing applications (M=3.78, SD =.655). The least competency level in this category is capturing pictures (M=3.76, SD =.732). Opening several programs simultaneously is the most prominent factor.

Table XVII displays the results on communication purposes using mobile technology among the respondent with

their friends. Three significant communication medium used as dimension in this subsection are: i) voice, ii) video and iii) email. Based on the above table, the results proved that the most skilful mobile technology among the students while communicating with their friends is email (M = 4.11, SD = .675), followed by voice and video with (M = 4.10, SD = .726) and (M = 4.09, SD = .699). Meanwhile, Table XVIII below present the results on competency of communication with their lecturers.

Based on Table XVIII, the output demonstrates that communicating with video is the highest competency level (M = 4.18, SD = .716), followed by communicating using email (M = 4.13, SD = .663), and communicating using voice (M = 3.98, SD = .665). The results for competency level with their family are presented in Table XVIII.

Table XIX displays the output of competency level of mobile technologies of communication factor. The most highly rank is communicating using email (M = 4.14, SD = .650), followed by communicating using voice (M = 4.12, SD = .737) and communicating using video (M = 4.05, SD = .668).

Table XX indicates the overall competency level of mobile technologies. The outcomes demonstrate that the competency with family is the highest (M = 12.3, SD = 1.56), followed by friends (M = 8.76, SD = 2.07), and lecturer (M = 2.96, SD = .798). Therefore, based on the results it is proven that students are more competent in the use of mobile technologies for communicating their families due to the distance between them and their families. It could be also speculated that they still have the influence of the surrounding environment that are skilful in the use of mobile technologies as a tool for social communication rather educational or study communication.

TABLE XVI. COMPETENCY OF MOBILE TECHNOLOGIES AS BASIC OPERATION TOOLS

Items	N	Min	Max	Mean	SD	Rank
Opening several programs simultaneously	148	3	5	4.20	.670	1
Organizing files into folders	148	2	5	3.97	.713	2
Recording videos	148	2	5	3.96	.737	3
Setting time and place for an event	148	2	5	3.95	.772	4
Sharing files and documents.	148	2	5	3.78	.787	5
Installing applications	148	2	5	3.78	.655	6
Capturing pictures	148	2	5	3.76	.732	7

TABLE XVII. COMPETENCY OF MOBILE TECHNOLOGIES FOR COMMUNICATION PURPOSES (FRIENDS)

Item	N	Min	Max	Mean	SD	Rank
Communicating using email	148	2	5	4.11	.675	1
Communicating using voice	148	2	5	4.10	.726	2
Communicating using video	148	2	5	4.09	.699	3

TABLE XVIII. COMPETENCY OF MOBILE TECHNOLOGIES FOR COMMUNICATION PURPOSES (LECTURERS)

Items	N	Min	Max	Mean	SD	Rank
Communicating using video	148	2	5	4.18	.716	1
Communicating using email	148	2	5	4.13	.663	2
Communicating using voice	148	2	5	3.98	.665	3

TABLE XIX. COMPETENCY OF MOBILE TECHNOLOGIES FOR COMMUNICATION PURPOSES (FAMILY MEMBERS)

Items	N	Min	Max	Mean	SD	Rank
Communicating using email	148	3	5	4.14	.650	1
Communicating using voice	148	2	5	4.12	.737	2
Communicating using video	148	2	5	4.05	.668	3

TABLE XX. RESULTS ON COMPETENCY OF COMMUNICATION WITH ALL

Items	N	Mean	SD	Rank
Family	148	12.3	1.56	1
Friends	148	8.76	2.07	2
Lecturer	148	2.96	.798	3

Table XXI presents the output of competency level of mobile technologies for collaboration purposes. Delivering presentation via video conference (M =3.04, SD =.880) is the highest competency level among the students, followed by attending a webinar (M =2.91, SD =.819) and sharing ideas via social media platform (M =2.81, SD = .732).

In Table XXII, it is stated above out of seven dimension tested, get the latest information by signing for newsletter (M =4.00 SD = .690) is the most prominent skills owned by the students. This is followed by acquiring the latest information via website (M =3.93 SD = .650), findings the current information by browsing (M =3.84 SD = .774), updating list of portal (M =3.77 SD = .681). Meanwhile, bookmarks portal and website related to the subject expertise is the next rated skills (M =3.76, SD = .846), followed by surfing the net to find references (M = 3.74, SD = .741) and lastly downloading from websites the learning materials (M =3.68, SD = .700).

Table XXIII presents the results on competency of digital citizenship among the undergraduates' students. The most highly rated skill is using other people's work lawfully (M =3.36, SD = .817). This is followed by learning knowledge related to cyber bullying (M =3.16, SD = .896), expressing self through media (M =3.10, SD = .831), creating password for privacy purposes (M =3.08, SD = .829), creating copyright of work (M =2.98, SD = .742). Meanwhile, finding information is the next competency level rated with (M =2.94, SD = .784), evaluating information (M =2.91, SD = .808), identifying valuable information (M =2.86, SD = .774). The least ranked competency skills among the students is using the information effectively (M =2.79, SD = .731).

Table XXIV demonstrates the findings on mobile technology competency of creativity and innovation tools. The most advanced skills owned by the undergraduates' students is uploading work (M =3.48, SD = .778). The second skill

mastered is creating an innovation /products using mobile technologies (M =3.47, SD = .812). Next competency ability is producing work (M =3.41, SD = .649), followed by performing advanced searches (M =3.40, SD = .687) and constructing an original work (M =3.35, SD = .648).

TABLE XXI. COMPETENCY OF MOBILE TECHNOLOGIES FOR COLLABORATION PURPOSES

Items	N	Min	Max	Mean	SD	Rank
Delivering a presentation via a video conference	148	1	5	3.04	.880	1
Attending a webinar	148	1	5	2.91	.819	2
Sharing ideas via other people's blogs, social forum, Social Networking Sites' walls etc	148	2	4	2.81	.732	3

TABLE XXII. COMPETENCY OF MOBILE TECHNOLOGIES FOR INFORMATION SEEKING TOOLS

Items	N	Min	Max	Mean	SD	Rank
Get latest information by signing up for newsletter of portals and websites	148	2	5	4.00	.690	1
Get latest information from subscribing portal/ website	148	2	5	3.93	.650	2
Finding latest information on subject / topic by browsing.	148	2	5	3.84	.774	3
Updating list of portals and websites in bookmark	148	2	5	3.77	.681	4
Bookmarking portals and website related to the field	148	1	5	3.76	.846	5
Browsing to find references	148	2	5	3.74	.741	6
Downloading learning materials from portals and websites	148	2	5	3.68	.700	7

TABLE XXIII. COMPETENCY OF MOBILE TECHNOLOGIES FOR DIGITAL CITIZENSHIP TOOLS

Items	N	Min	Max	Mean	SD	Rank
Using other people' works lawfully	148	2	5	3.36	.817	1
Learning what to do if involved in cyber bullying & how to overcome	148	2	5	3.16	.896	2
Expressing myself through digital media	148	2	5	3.10	.831	3
Creating strong passwords to protect my private information	148	2	5	3.08	.829	4
Protecting my digital works through the copyright	148	1	5	2.98	.742	5
Finding out required information	148	1	5	2.94	.784	6
Evaluating information	148	1	5	2.91	.808	7
Identifying valuable information	148	1	5	2.86	.774	8
Using information effectively	148	1	5	2.79	.731	9

TABLE XXIV. RESULTS OF MOBILE TECHNOLOGIES AS CREATIVITY AND INNOVATION TOOLS

Items	N	Min	Max	Mean	SD	Rank
Uploading work	148	2	5	3.48	.778	1
Creating an innovation/product	148	2	5	3.47	.812	2
Producing work	148	2	5	3.41	.649	3
Performing advanced searches	148	2	5	3.40	.687	4
Constructing an original work	148	2	5	3.35	.648	5

IV. Conclusions and Implications

The focus of this study was to report the degree of utilization among first year undergraduate students when using mobile technologies, the level of competency among first year undergraduate students when using mobile technologies, the level of basic operation, communication and collaboration, information Seeking, digital citizenship and creativity and innovation skills in mobile technologies competency separately. The target sample of this study consists of 148 freshmen students majoring in computer science and information technology at Hadhramout University on their level of frequency use of these technologies. This study is unique, and it adds to the body of the research. The uniqueness of this study is due to the few searched topics in this field in the context (Hadhramout). The instrument used to collect the data of the study is adopted from [12] which is quoted from [7] [10] which has been adapted to suit the research objectives.

This study finds that undergraduate students in faculty of computer science and information technology are highly utilizing mobile technologies (M = 151, SD =12.1). Similarly, they have good capabilities and skills in using them (M= 141, SD= 13.5). The results also uncovered that communication is highly used by these students (M =41.4, SD = 2.62) followed by basic operations (M =28.3, SD = 4.02). These results of utilization are supported by students' competency in using the same purposes where it is discovered that students are so competent in using mobile technologies for communication (M =33.4, SD = 378). Moreover, it is revealed that students' skilfulness in basic operation is also significant (M =27.4, SD = 3.49).

Through results which shows the second most used purpose is information seeking, it could be implied that students become more aware about getting what will enrich their backgrounds with knowledge in their field. Thus, there must be much focus paid to this method to have better level of students. To have this fulfilled, it is advised that faculty of computer science and information technology to strengthen their bandwidth and permit students to have more access to the internet to encourage them using mobile technologies for that purpose. Moreover, to have a better engagement of students, it is desirable to activate the university application. The use of this application may guide students' choice of information to what is more beneficial and related to their study. Furthermore, activating the app may also elevate the effect of competency in communication uses in which students will be forced to contact only lecturers or classmates. This will result in more engagement and directed communication that will lead to better learning and utilization of mobile technologies.

Despite the significant competency of CS and IT students in basic operation and information seeking, it is preferable that these students acquire more competencies to have higher level of mobile technologies utilization. The author in [8] who declared the importance of students possessing technical skills will assist them in their workforce when graduating, support this assumption. Moreover, they detailed that these skills must be gained prior entering working places, which means HE institutions must empower their students with these skills. This is achievable through training teaching staff on how to integrate technology into classes. There must be also workshops to provide them with techniques that will enable them directing these learners after arming them with needed capabilities to the desirable uses that will strengthen their skills and make them ready for the work force. This is in a line with [9] which states that education institutions must link their teaching staff to chance that will provide them with technological skills enabling them to have better adoption of continues emerged technologies effectively.

As it can be seen from results that HE undergraduates are utilizing the gadgets for communicating with their friends or family, which means social type of use; therefore, lecturers ought to shift these learners to communicate more for the sake of learning through contacting their lecturers and peers in the campus and other countries. Lecturers must open channels for students to show them how communication can be a good tool with the existence of internet and these mobile technologies. They may innovate through trying the flipping classroom or problem based or any new method of teaching since mobile learning is facilitating these methods a lot.

It is highly recommended since this college is adopting mobile learning method to activate the university app to be a source of knowledge and learning. It is recommended to get use of the app to contain an LMS (learning management system) that will not only help students to register and know their subjects but also communicate and collaborate in different college activities that happened digitally. Moreover, the experience of using M-Learning in this college must be expanded to other colleges to show them the benefits the faculty in outcomes results and learning quality as well.

It is also recommended to conduct this research in another governorate where it might be found more colleges are applying this method. This change will result in having different results since the participants will be greater and the context varied a little bit. It is also recommended to have this study done in qualitative mode where the researcher can get more in depth data about why students choose certain purpose rather the other. With the existence of the app and the website, it is highly recommended to conduct and exploratory study on students and teachers' perspectives towards the existence of these tools in the campus. This study will be like an evaluative study covering strengths and weakness of both tools. Studies must be conducted to check teachers' point of view towards this new method along with whether the infrastructure of the college is supportive or not.

Acknowledgment

This publication was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia.

Reference

[1] Barbaux, M. T. (2006). From lifelong learning to m-learning. *Association for Learning Technology*, 132.

[2] Corbeil, J. R., & Valdes-Corbeil, M. E. (2007). Are you ready for mobile learning?. Educause Quarterly, 30(2), 51.

[3] Draugalis, J. R., Coons, S. J., & Plaza, C. M. (2008). Best practices for survey research reports: a synopsis for authors and reviewers. American journal of pharmaceutical education, 72(1), 11.

[4] Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). Internal validity. How to Design and Evaluate Research in Education. New York: McGraw-Hill, 166-83.
[5] Keller, J. L. (2011). Advancing student success with competency points: Elevating engagement and motivation in community college English composition students. Community College Journal of Research and Practice, 35(6), 484-504.
[6] Lei, J. (2010). Conditions for ubiquitous computing: What can be learned from a longitudinal study. Computers in the Schools, 27(1), 35-53.
[7] Mather, C. A., Marlow, A. H., & Cummings, E. A. (2013). Using Web 2.0 to support Continuing Professional Development of health professionals: Acting locally, engaging globally. In Teaching Matters: Open UTAS to the World (p. 24)\.
[8] Matias, A., & Wolf, D. F. (2013). Engaging students in online courses through the use of mobile technology. In Increasing Student Engagement and Retention Using Mobile Applications: Smartphones, Skype and Texting Technologies(pp. 115-142). Emerald Group Publishing Limited.
[9] Oblinger, D. G. (2010). a Commitment to Learning: attention, engagement, and the Next generation. Educause Review, 45(5), 4.
[10] Papoutsi, C., & Drigas, A. S. (2017). Empathy and Mobile Applications. International Journal of Interactive Mobile Technologies, 11(3).
[11] Sedek, M., Mahmud, R., Jalil, H. A., & Daud, S. M. (2014). Factors influencing ubiquitous technology usage among engineering undergraduates: a confirmatory factor analysis. Middle-East Journal of Scientific Research, 19, 18-27.
[12] Severino, S., & Messina, R. (2010). Analysis of similarities and differences between on-line and face-to-face learning group dynamics. World Journal on Educational Technology, 2(2), 124-141.
[13] Traxler, J. (2007). Defining, Discussing and Evaluating Mobile Learning: The moving finger writes and having writ.... The International Review of Research in Open and Distributed Learning, 8(2).
[14] Tuparov, G., Alsabri, A. A. A., & Tuparova, D. (2015, November). Students' readiness for mobile learning in Republic of Yemen—A pilot study. In Interactive Mobile Communication Technologies and Learning (IMCL), 2015 International Conference on (pp. 190-194). IEEE.
[15] World Bank Group. (2016). World development report 2016: digital dividends. World Bank Publications.
[16] Oinas-Kukkonen, H., & Harjumaa, M. (2008, June). A systematic framework for designing and evaluating persuasive systems. In International conference on persuasive technology (pp. 164-176). Springer, Berlin, Heidelberg.
[17] Liaw, S. S., Hatala, M., & Huang, H. M. (2010). Investigating acceptance toward mobile learning to assist individual knowledge management: Based on activity theory approach. Computers & Education, 54(2), 446-454.

Infrastructure Study for Solving Connectivity Problems Through the Nile River

(A Case Study on the Assuit-Delta Reach)

Noha Kamal[1]
Associate Professor
Nile Research Institute (NRI)
National Water Research Center (NWRC), Cairo, Egypt

Ibrahim Gomaa[2]*
Assistant Professor
Computers and Systems Department,
National Telecommunication Institute (NTI), Cairo, Egypt

***Abstract*—Fiber optics cables present various benefits over regular cables when used as a data transportation medium in today's communication networks. It is noted that there are significant challenges in the connectivity of inner cities that are located far inland away from the coastal areas. Most of the networks developed in Africa, especially in Egypt, are connected via submarine cables flowing across coastal areas. Very few connections are constructed to connect inner cities by crossing the Nile. The Nile River is characterized by a wide area, offering a natural path for underwater cables' laying areas. In this study, the analysis and evaluation of the laying of these cables along the bed of the Nile River in Egypt, rather than crossing it, is investigated. There are many issues with laying fiber optic cables across the Nile River. Some of these are the requirement of using more than one node over fiber optic cable for each. When the number of nodes increases, the cost of installation and drilling effort increases with each node. The fiber optic cable path along the Nile River is simulated with a numerical model (Delft-3D). Two different scenarios for laying cables were applied and analyzed to evaluate the effect of the predicted water surface and sediment profiles on the fiber optic cable path. Based on the results obtained, the fiber-optic network infrastructure is proposed to solve connectivity problems by laying fiber optic cables along the Nile River.**

Keywords*—Communications; optical fiber cables; delft3d; underwater / river crossing cable

I. Introduction

Communication is very important in the life of human beings since ancient times. Underwater optical fiber communication technologies present new advancements in this field. Underwater optical fiber technology has evolved over the years and is growing at high rates [9]. Fiber-optic usage in communication networks has many advantages and disadvantages compared to electric wiring cables, especially with long distances. Fiber optics have various benefits such as high data security (since there is no electromagnetic radiation from cables), immunity to electromagnetic interference, lack of current-induced sparks risks (having no conducting current used), using small and lightweight materials, and high operating bandwidth over long distances [11]. On the other hand, these cables have some disadvantages such as cost because cables have expensive installation although they last longer than copper cables. Also, optical fibers require repeating at distance intervals, and Fragility [21] [8]. In addition to that, optical fibers require more protection around cables compared to copper [1]. Various projects aim to link Africa to the world networks through undersea fiber-optic cable networks; as shown in Fig. 1 [20]. However, there is a difficult challenge in the connectivity of inner towns and cities that are located far inland away from the coastal areas.

In Egypt; the Abu Talat city which is marked by number 1 is linked to the undersea Cables Europe India Gateway (EIG), Middle East North Africa (MENA) Cable System/Gulf Bridge International, TE North/TGN- Eurasia/SEACOM/Alexandros network systems. Alexandria city no.2 which is connected to Cables Aletar, FLAG Europe- Asia (FEA), Hawk, IMEWE, SeaMeWe-3, and SeaMeWe-4. Suez city no.3 connected to cables FALCON, FLAG Europe- Asia (FEA), IMEWE, SeaMeWe-3, and SeaMeWe-4. Zafarana city no.4 connected to cables Europe India Gateway (EIG), Middle East North Africa (MENA), Cable System/Gulf Bridge International, SEACOM/Tata TGN-Eurasia. [20].

The objectives of this research are to study, analyze, and evaluate the laying fiber optic cables system along the bed of the Nile River in Egypt. Because of the fiber optic network characteristics and advantages, fiber-optic cables are considered an effective alternative to consider for a new communication network in long and spacious lands such as in Egypt. This is in comparison to using either satellite for telecommunication or laying copper wire coaxial cables for long distances [18].

Fig. 1. Map of Africa with Coastal undersea Fiber-Optic Networks. Egypt Cable Locations are Encircled. [20].

*Corresponding Author

II. Related Works

Several researchers have provided a detailed analysis of submarine optical fiber cable transmission systems. Ammar A. Saleh, et al. [15] study and analyze a proposal suitable for African Nile River basin countries to lay new fiber optics cable networks submerged under the river's water level. This proposal discusses the challenges that are faced by common ground cable networks. Paul D. et al. [3] provided an analysis of undersea bed fiber optic cable network that is reported on several cable networks TAT, their reliability, and cost considerations. Recently, Navneet A. and Ajay V. [12] Neal S. Bergano [1], investigated undersea fiber optic cables that make the web worldwide. Modern cable systems installed up to last year are capable of transmitting about 1000 Gbps over each fiber pair [4]. In [7] they studied newly retiring fiber optic telecommunication cables, which offer far greater opportunities for the scientific community. Hsiang et al. [13] focused their work on the research and development of an integrated underwater environmental monitoring system, which is planned to be applied on the offshore wind farm of Taiwan. In [6] SAICi/MariPro was contracted by Alcatel TCC in Australia, to install a 240-kilometer long submarine fiber optic cable system across the Bass Strait in August 1995. This eighteen fiber cable system provided telecommunication services between mainland Australia and Tasmania for customers of TELSTKA. Jeffcoat et al. [2] introduced considerations in the design and laying of a new undersea fiber optic cable system. Jurdana et al. [14] analyzed underwater fiber-optic cables, evaluated their impact on the marine environment, and addressed possible threats to submarine cables from human activities and natural hazards. This work provided guidelines for improving the installation of cables, and also their security. Msongaleli et al. [17] inspected the disaster-aware submarine fiber-optic cable deployment process problem to minimize such expected extra costs in case of a disaster. Namihira et al. [5] studied the optical fiber submarine cables Polarization fluctuation characteristics under 8000-m deep-sea environmental conditions, optical fiber submarine cable conjunction under periodic variable tension, and the performance of cables through and after installation.

III. Study Area Description and Data Collection

To achieve the objectives of this study, the following methodology is applied as shown in Fig. 2.

1) Develop a database for fiber optic cables that cross the Nile River.

2) Determine the proposed fiber optic cable longitudinal path along Nile River using recent hydrographic survey data in 2016.

3) Use a Delft-3D numerical model to simulate two different scenarios for laying the fiber optic cable along the bed of Nile River by applying maximum and minimum flow, predict the morphological changes from the year 2016 to the year 2030, which could affect the efficiency of the fiber optic cables.

4) Evaluate and compare two different scenarios of laying fiber optic cables along the study's reach.

5) Evaluate the economic and environmental impacts of the ground cables and submerged cables under the Nile River.

6) Propose Nile River fiber-optic network infrastructure to solve the connectivity problems.

A. Study Area Description

The Nile River in Egypt mainly consists of a long single-channel followed by two branches forming the Delta. The single channel's length is 953.5 km downstream Old Aswan Dam (OAD) in the south to just upstream Delta Barrages north of Cairo. The river is divided into four reaches. This paper focuses on the fourth reach as shown in Fig. 3, which is the longest and extends between Assiut and Delta Barrages covering a total distance of 408.75 km. It is marked by having numerous natural phenomena represented in many islands' characteristics, and many bends. This study focuses on a segment from km 918 upstream Delta Barrage to km 937 downstream OAD.

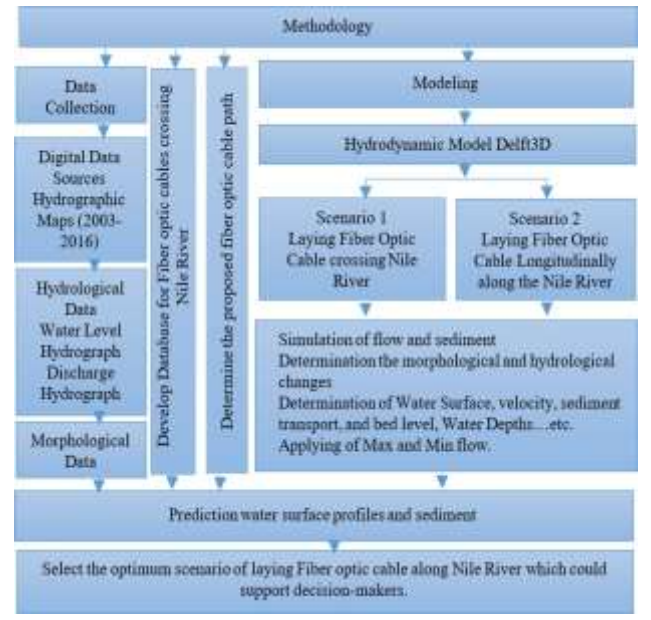


Fig. 2. Research Methodology.

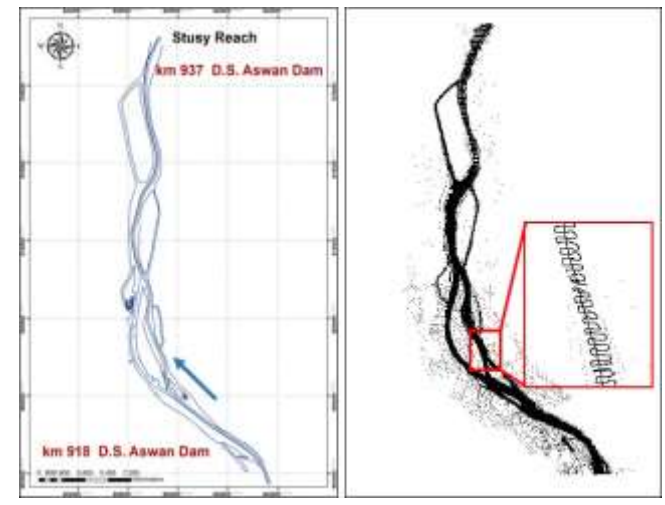


Fig. 3. Reach Location and Bathymetry Data of Study Area.

B. Data Collection

1) Hydrographic Data: 20 km of the Nile River was hydrographically surveyed by the Nile Research Institute (NRI) in 2003, and 2016 respectively to reveal the riverbed morphology.

2) Hydraulic Data: the measurements of velocity were collected at the same time as the survey in 2003, and 2016. Cross-sections were selected to overlay the length of the study area. At the soil laboratory of NRI, the samples were analyzed. The results of this analysis indicated that the bed material consisted of sand and silt. The sand ranges between 95.44% and 99.07% of the bed material sample. As for the silt, it varies in the range between 0.18% and 4.43% of the bed material sample.

3) Hydrological Data: Hydrologic data is needed to initiate the model's boundary conditions. Besides, the discharge variation released through the river and its corresponding fluctuations cause a sediment transport process, which in turn, causes morphological changes through the river bed. The monthly average discharges of three years that were released downstream Assiut barrages, and the corresponding monthly average water levels have been measured at Assiut gauge station were gathered from the historical records from 2000 to 2009 as shown in Fig. 4. These three years were selected as they were distinguished by maximum and minimum flow releases respectively to represent the most critical situation that can affect the morphological characteristics of the river.

The maximum monthly average in three years of the water level and discharge of historical data from (2000-2002) and the minimum monthly average in three years of the water level and discharge of historical data from (2003-2005) are shown in Fig. 5a and 5b, and Fig. 6a and 6b.

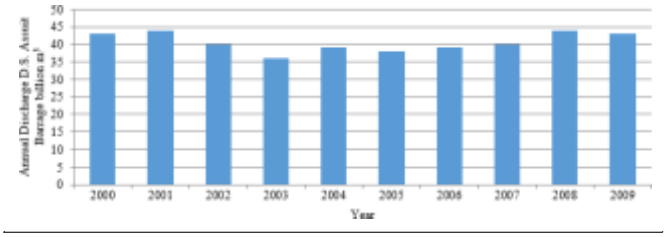
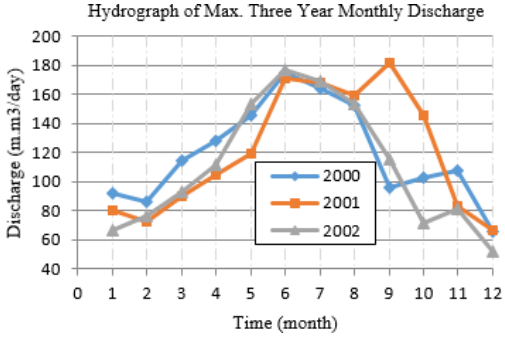
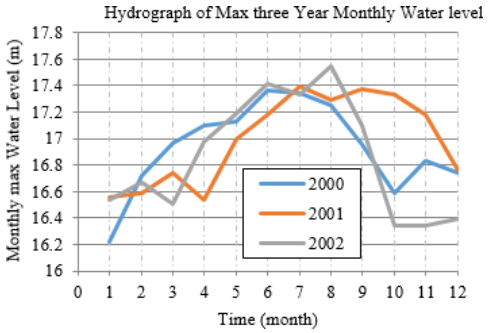


Fig. 4. River Discharge at Study Area.

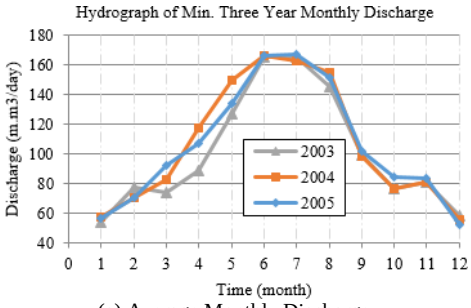


(a) Average Monthly Discharge.

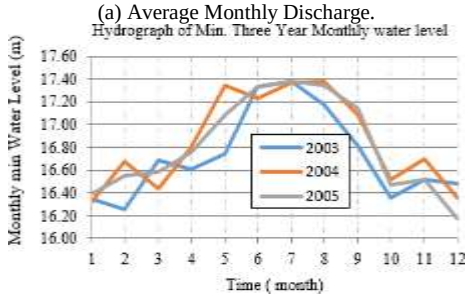


(b) Average Monthly Water Level.

Fig. 5. Nile River Hydrograph of Maximum Three Year (2000- 2002).



(a) Average Monthly Discharge.



(b) Average Monthly Water Level.

Fig. 6. Nile River Hydrograph of Minimum Three Year (2003- 2005).

IV. Developing a Database for Fiber Optic Cables Crossing the Nile River

In this part, the development of a database is necessary and required to draw a complete picture with up-to-date information on the number and locations of fiber optic cables crossing the Nile River in Egypt. Fig. 7 shows the flexible interface of the developed database by using visual basic programming language and Geographic Information system GIS. This database stores information about all the cables.

Fig. 7. Developed Database of Fiber Optic Cables Crossing Nile River.

V. Determine the Proposed Longitudinal Fiber-Optic Cable Path along Nile River

By using NRI's recent hydrographic survey data from 2016, as well as the navigational path, longitudinal fiber optic cable path can be determined according to the deepest points as shown in Fig. 8. Furthermore, the calibration of that proposed path was applied by using the model shown in Fig. 9.

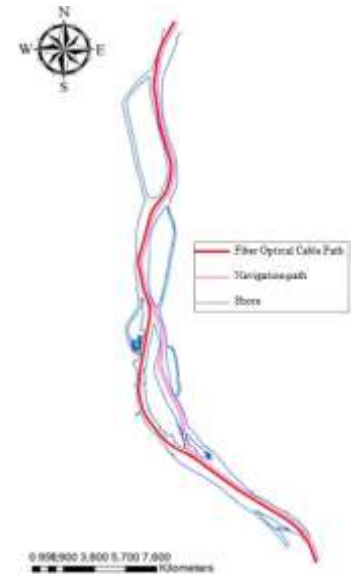


Fig. 8. Proposed Fiber Optical and Navigation Paths.

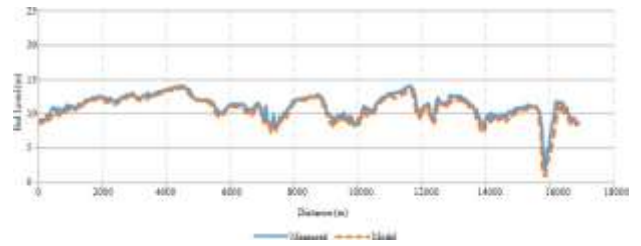


Fig. 9. Bed Level Calibration Proposed Fiber-Optic Path Longitudinal Section.

VI. Numerical Model and Simulation

A. Numerical Model

Delft3D is the hydrodynamic module of Delft3D, which is an integrated program for modeling water flows, waves, water quality, particle tracking, ecology, sediment, and chemical transports and morphology [19]. The Meyer-Peter-Muller sediment transport relation will be used [12].

1) Model preparation and grid generation: Delft3D model generated a grid network. This grid covers a distance of 18 km along the shoreline. A fine grid (5m*5m) used in the model as shown in Fig. 10. The initial boundary condition is defined as the initial water levels. The initial water levels were used to simulate the flow characteristics. Both upstream and downstream boundary conditions were given to the model as inputs. The upstream boundary condition was the discharge downstream from the Assiut barrage. Also, Upstream Delta Barrage water levels were used as the downstream boundary condition.

2) Model calibration: The model calibration process was done for flow velocity distribution and also sediment transport. It can be noted that there are a sound and logical agreement between the computed and measured values of the flow velocity distributions and morphological changes at the chosen cross-sections. The model was carried out by adjusting roughness coefficients at various locations along the modeled study reach, the calibration process was run to attain the finest agreement between measured and resulted values of the model. Fig. 11 shows the velocity calibration process and velocities comparison at different cross-sections along the study area. Besides, Root Mean Square Error (RMSE) which is presented in Eq. 1 was computed to quantify the model's performance for the observed and measured bed level values [10].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}} \quad (1)$$

Where X_{obs} is observed values and X_{model} is modeled values at time/place.

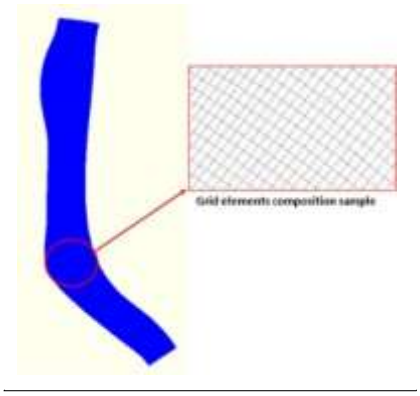


Fig. 10. Studied Reach Grid.

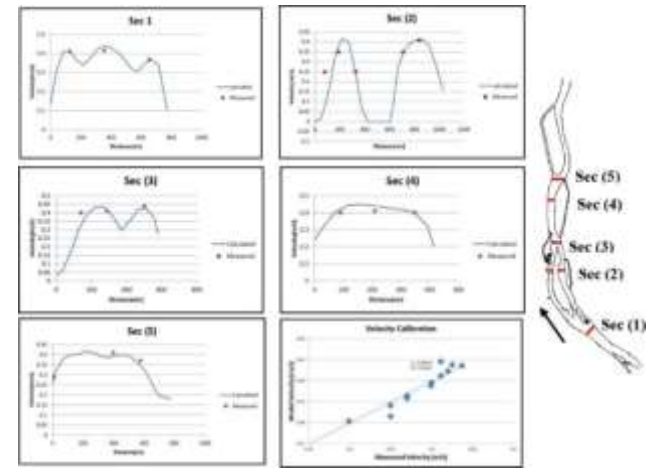


Fig. 11. The Velocity Calibration.

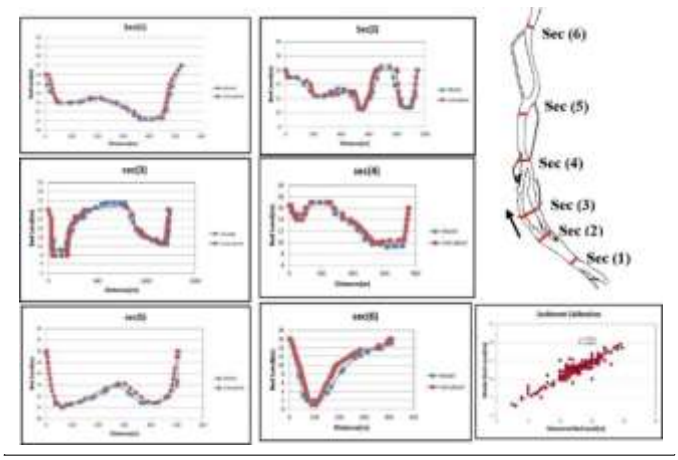


Fig. 12. The RMSE for the Observed and Measured Sediment Values.

Fig. 12 shows the calibration of the bed level and the comparison of observed and measured cross-sections through the area of study. Also, RMSE which was presented by Eq. 1 is used to quantify the model performance for the observed and measured bed level values.

3) Model simulation: The model simulated two different scenarios to achieve the main objectives of this paper. In the model simulation, the recent hydrographic survey data from the year 2016 is used to predict the morphological changes for two different scenarios of fiber optical cable path to select the optimum scenario.

a) The first scenario, fiber optical cable crossing the Nile River.

b) The second scenario is that the fiber optical cable is laid and installed along the Nile's riverbed.

These scenarios are simulated by Delft3D by applying maximum and minimum flows.

1) Scenario 1: Fiber optical cables that cross the Nile River according to our database, there are twenty-five fiber optic cables that cross the Nile River as recent as 2019. In this scenario; the Delft3D model simulation for selected fiber optic cable crossing the Nile River in the study reach is represented. The selected fiber optic cable crosses the Nile at km 928.15 from Aswan High Dam in May 2007 as shown in Fig. 13. Also, the natural changes of the river that could affect the cable efficiency have been evaluated. The numerical simulation was performed for maximum and minimum flows at the study reach to indicate the locations of deposition and erosion. According to NRI technical reports; as this cable is laid directly on the riverbed, it falls 25 cm below the bed as shown in Fig. 14.

Fig. 15 shows the predicated water depths, water velocities, and bed levels patterns for the initial river bed levels surveyed in 2016 and the predicted for the year 2030 in case of releasing discharges for maximum and minimum. As shown in Fig. 15a, as a result, after running the model for fourteen years, it is clear that the maximum erosion occurs when the maximum discharge is passed compared to minimum discharge, particularly at the location of the fiber cable that crosses the Nile. Furthermore, Fig. 15b, and 15c, the initial velocity distribution in 2016 ranges between 0.4 and 0.5 (m/s).

The predicted velocity distribution in 2030 is high in the maximum case at many locations, particularly in outer curves regions, at the location of fiber cable, which ranges between 1.0 and 1.2 (m/s). In the same pattern, the velocity distribution at the minimum case was introduced. It can be noted that in the case of maximum flow, erosion has a significant effect on the location of fiber cable path that crosses the Nile than sedimentation. As a result, this will affect the fiber optic cable path which crosses the Nile, especially in the case of maximum flow.

2) Scenario 2: Laying fiber optic cable along the riverbed. The model simulated laying the fiber optic cable along the bed of Nile River as shown in Fig. 16. In this scenario, a numerical simulation was performed for maximum and minimum flows at the upstream and water levels at the downstream boundary of the study reach.

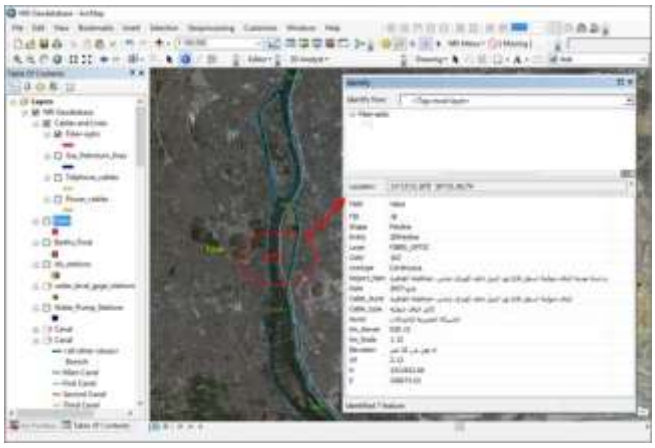


Fig. 13. Fiber Optic Cables Crossing Nile River.

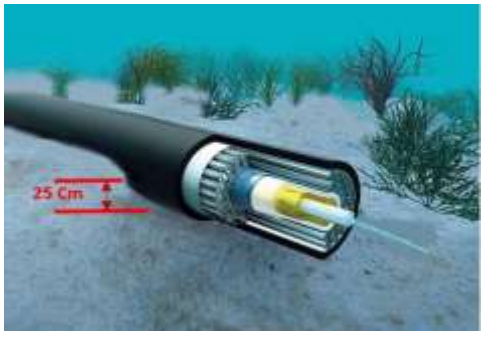
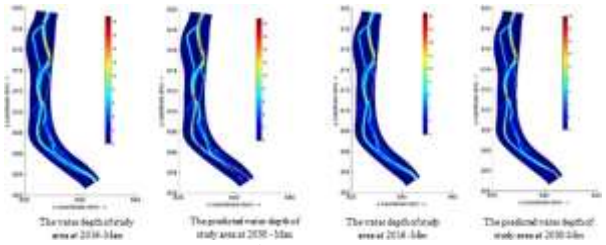
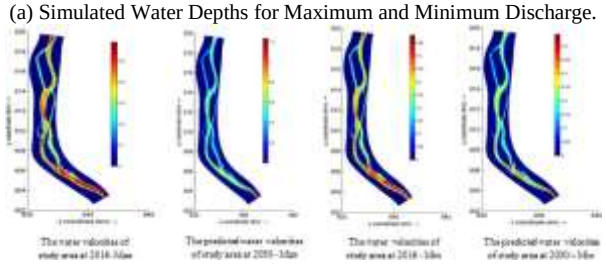


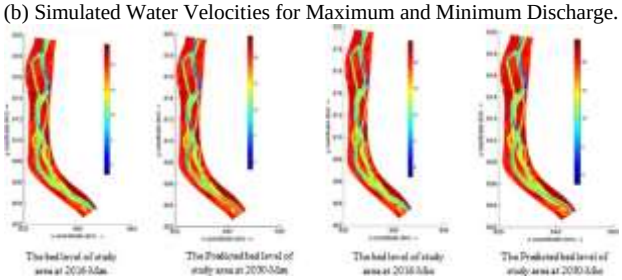
Fig. 14. Laying Fiber Optic Cable Directly on the Bed Level of Nile River.



(a) Simulated Water Depths for Maximum and Minimum Discharge.

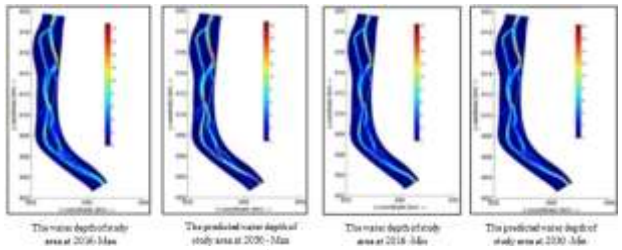


(b) Simulated Water Velocities for Maximum and Minimum Discharge.

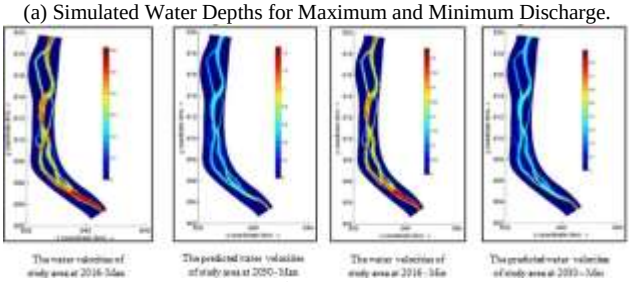


(c) Simulated Bed Levels for Maximum and Minimum Discharge.

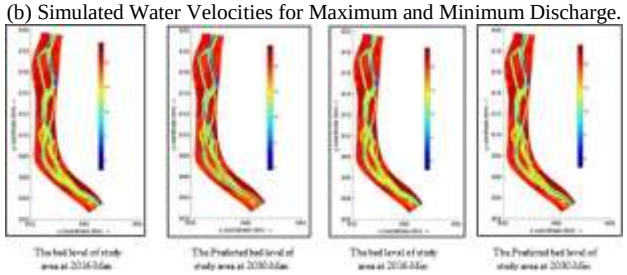
Fig. 15. Simulated Water Depths, Water Velocities, and Bed Levels Patterns for Scenario1.



(a) Simulated Water Depths for Maximum and Minimum Discharge.



(b) Simulated Water Velocities for Maximum and Minimum Discharge.



(c) Simulated Bed Levels for Maximum and Minimum Discharge.

Fig. 16. Simulated Water Depths, Water Velocities, and Bed Levels Patterns for Scenario 2.

VII. Results and Discussions

The proposed fiber optic cable path is affected by certain factors related to river processes as maximum and minimum flow, erosion, and deposition. Hence, it is very important to simulate and check with different flow cases. Any changes in river morphology are prohibited unless they are of absolute importance such as maintenance operations. That is including pump intakes and pipeline crossing under the river. Fig. 17 shows the predicted max and min flow morphological changes in 2030 for the first scenario, where the fiber optic cable was laid by crossing Nile River at km 928.15 downstream OAD through the tunnel drilling of the river for laying the fiber optic cable along the riverbed by 3 (m). It can be noted that the rate of erosion was greater than deposition, which could affect the safety and efficiency of the fiber optic cable.

On the other hand, Fig. 18 shows the longitudinal bed level profiles for both the initial riverbed in 2016 and the predicted in 2030 for the release of maximum and minimum flows for the second scenario, where the fiber optic cable is laid along the Nile River bed. It can be observed that in the first scenario, the erosion rate relative to the length of the cross-section area is greater in the second scenario. Also, in the first scenario, the deposition rate relative to the length of the cross-section is greater in the second scenario. It is clear that in scenario 1, the efficiency of the fiber optic cable path will be affected by the high erosion rate compared to the second scenario. Hence, it can be concluded that the proposed fiber-optic cab1e laying scenario 2 is better than scenario 1.

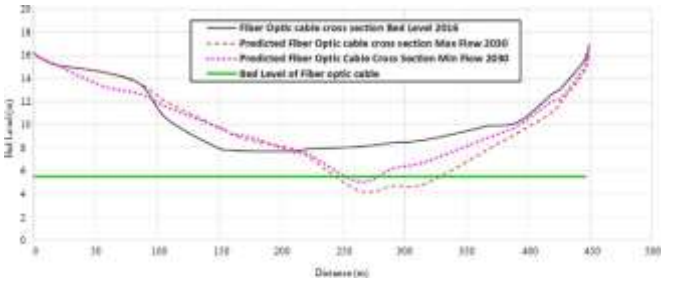


Fig. 17. Predicted Morphological Changes at Max and Min Flow (Scenario 1).

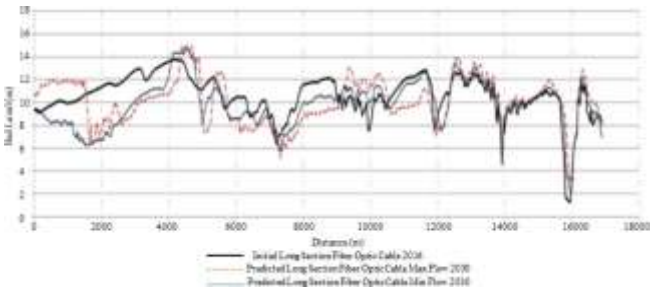


Fig. 18. Predicted Morphological Changes at Max and Min Flow (Scenario 2).

VIII. Evaluating and Comparing the Economic, Environmental Impacts of the Ground Cables and Submerged Cables under the Nile River

A comparison is introduced detailing the use of submerged river cab1es versus ground cables. It has been demonstrated that this type of network can be used with significant advantages over ground cable networks, especially for countries with long river lengths. While most Egyptian cities and towns are located deep inland to the west and south of the Mediterranean and the Red Sea coasts. There are many challenges facing ground fiber optic cable networks in Egypt. In particular, they involve high-cost drilling operations, especially when connecting cables to remote locations. Furthermore, they may suffer from loss of cables and equipment (e.g.: due to theft and vandalism). Besides, it is time-consuming and maintenance is costly. Moreover, there is a need for repeaters grid and Land permit costs for areas on which cable network is established. Table I shows these evaluations of economic and environmental impacts based on running costs, maintenance, repeaters, and power. According to this evaluation, a suitable proposal will be presented for the establishment of new fiber optic cable networks in Egypt submerged under the Nile River. Such networks could be considered as the core network of African countries especially those which are part of the Ni1e River basin.

TABLE I. Economic and Environmental Impacts Comparison between the Ground and Submerged Cables under River Nile

Economic/environmental Impact Factor	Ground Cables Network	Underwater Cables Network
Repeaters	Repeater grid	Repeater along River
Power	Power grid	Power along River
Theft	Frequently	Hardly
Maintenance	Time-consuming and costly	rarely
Security	Vulnerable	impervious
Damage	vulnerable	impervious
Human interventions	Frequent	Negligible
Monitoring points	Difficult	Easy
General Costs	High	Standard
Land Cost	High	Standard
Running cost	High	Standard
Service Continuity	Low	High
Materials	Standard	Standard
Wars	Direct Effect	Negligible
Temperature	High Effect	Low Effect
Medium	ground	water
Earthquakes	High Effect	Low Effect

IX. Proposed Nile River Fiber Optic Cable Features

A. Fiber Optic Cables

Fiber Optic cabling has revolutionized the way that data travels globally. These cables are characterized by high- speed, which transmits data at much faster rates without any quality degradation compared to copper wires. Also, they contain an outer optical casing, which surrounds the light. The core can be configured in two different types:

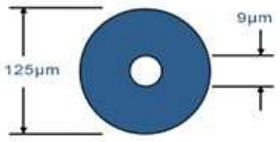
1) Single-mode fiber optic cable: This type has a small core size (less than 10 µm) as shown in Fig. 19a. It allows the transmission of one light ray only. So, as the light passes through the single-mode fiber core, a little reflection of light is formed. This decreases fiber attenuation and further increases the signal's ability to travel. Thus, single-mode fibers are usually used in long distances up to 1000 km and high bandwidth applications.

2) Multimode fiber optic cable: Multimode fibers are characterized by larger cores (62.5 μm or 50 μm) - as shown in Fig. 19b - which introduce more data. This will allow more light reflections, and increase the rate of dispersion and attenuation, which may reduce the signal quality over long distances. Therefore, it can be noted that in short distances, multimode fibers are often preferred.

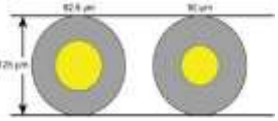
B. Recommended Fiber Optic Cable

According to the National Telecommunication Institute (NTI), Egypt, reports and recommendations, the recommended type of fiber optic cable (Fig. 20) Underwater / River crossing type optical fiber cable, based on ITU-T recommendations, IEC 60794, IEC 60332, IEC 304, or EIA/TIA-598 standards. This type of optical cable is characterized to be suitable for the transmission of voice, data, and broadband over long distances, and local area networks with installation methods underwater for river crossing [16].

The type of these cables marked by MLT (Multi-Loose Tube) design, high fiber capacities, Single-mode, multimode (50/125, 62.5/125), waterproof dry core design, high tensile strength, improved compressive strength, bituminized polypropylene yarns as an outer jacket.



(a) Single-Mode Fiber-Optic Cable.



(b) Multi-Mode Fiber-Optic Cable.

Fig. 19. Fiber Optic Cable Types.

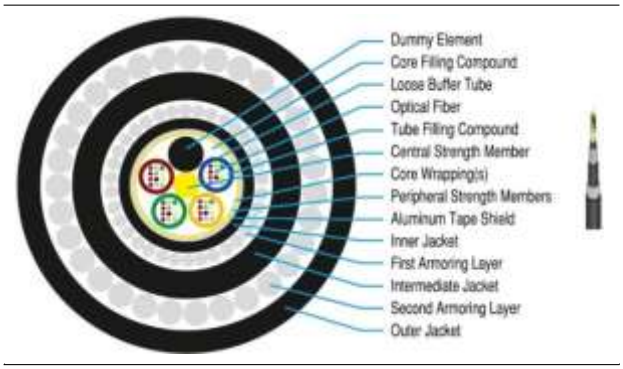


Fig. 20. Proposed Underwater Optical Fiber Cable.

Based on the results and discussions of the Delft3D model, it can be concluded that it would be beneficial to consider constructing a fiber optic cable network along the bed of the Nile River to provide connectivity for inland cities. This infrastructure should be constructed as follows:

1) Installation of monitoring and control centers along the Nile River path for every 1000km. This distance is the longest distance that can be traveled by light through these cables before repeaters are required. Since the Nile River at the Egyptian border has a total length of 1440 km, one repeater will be needed.

2) Installation of control centers can serve as checkpoints for other essential services such as water level measurements, water quality monitoring, and other public safety services.

3) Based on results and discussion; it is recommended that the second scenario be used to lay a fiber cable network along the riverbed in Egypt.

4) According to the recommendations of NTI, the single-mode fiber optic cable type (underwater / River crossing type optical fiber cable) should be used.

X. Conclusion and Recommendations

It was noticed that the majority of fiber optic cable networks are established via submarine cables through the coastal areas. Constructed connections are not enough to connect inland towns and cities. This paper introduces a proposal to utilize the natural path of the Nile River in Egypt to construct a fiber optic cable system submerged under the river's water. This would provide coverage to inland areas that are located closer to riverbanks. Delft3D hydrodynamic model was used to simulate two different scenarios for laying fiber optic cable submerged under a riverbed in Egypt. Based on the results and discussions, it can be observed that in the first scenario, the erosion rate was greater than the deposition rate at the location of fiber optic cable compared to the second scenario. This could harm the safety and efficacy of the fiber optic cable. Furthermore, there are many issues with laying fiber optic cables across the Nile River. Several of these are the requirement of using more than one node over fiber optic cable for each. When the number of nodes increases, the cost of installation and drilling effort increases with each node. It can be concluded that scenario 2 is more suitable than scenario 1 for laying such cables along the Nile River in Egypt. Consequently, we found that it would be beneficial to consider constructing and installing a fiber optic cable network along the Nile riverbed to provide connectivity to inland cities. In this study, it is strongly recommended that a n action plan be developed for the proposed infrastructure as a fiber optic system along the Nile River. Finally, it is recommended to take into consideration the coordination with the affiliated authorities of public facilities and services that pass through the Nile River in Egypt to obtain all information about the locations and paths of these services such as gas lines, electricity, and phone cables.

Acknowledgments

The authors thank Shimaa Badr, National Telecommunication Institute, Egypt for technical support.

References

[1] Neal S Bergano. "Undersea fiber-optic cables make the web worldwide". In: 17th Opto- Electronics and Communications Conference (OECC 2012) Technical Digest. Busan, Korea, July 2012.

[2] C Jeffcoat, J Jackson, and J Ewald. "Considerations in Design and Laying a New Undersea Fiber Optic Cable." In: Challenges of Our Changing Global Environment. OCEANS '84 MTS/IEEE, 1984. DOI: DOI: 10. 1109 / oceans.1984.1152235.28.

[3] D Paul et al. "Undersea fiber optic cable communications system of the future: Operational, reliability, and systems considerations". In: IEEE journal of light technology 2.4 (1984), pp. 414–425.

[4] Y Iwamoto and H Fukinuki. "Recent advances in submarine optical fiber cable transmission systems in Japan". In IEEE Journal of Lightwave Technology 3.5 (Oct. 1985), pp. 1005–1016.

[5] Y Namihira et al. "Dynamic polarization fluctuation characteristics of optical fiber submarine cables under various environmental conditions." In: Journal of Lightwave Technology 6(5) (1988), pp. 728–738. DOI: DOI: 10.1109/50.4059.

[6] M R Harrison. "Installation and burial of a 240-kilometer long fiber optic telecommunication cable in the Bass Strait using the "ship-of-opportunity" installation technique". In: Challenges of Our Changing Global Environment. OCEANS '95 MTS/IEEE, 1995. DOI: DOI: 10.1109/oceans.1995.526760.

[7] R Butler. "Scientific re-use of retired undersea fiber optic telecommunications cables." In: International Conference Physics and Control. (Cat. No.03EX708), 2003. DOI: 10.1109/SSC.2003.1224153.

[8] Amir Hosseini, M K Ramezani, and M Banae. "Submarine cable installation between production platform and satellite wellhead platform of south pars gas field - phase 1 in the Persian Gulf". In Challenges of Our Changing Global Environment. OCEANS '04 MTS/IEEE, 2004. DOI: DOI: 10. 1109 / oceans. 2004. 1402923.

[9] M P Wood and L Carter. "Whale Entanglements with Submarine Telecommunication Cables." In: IEEE Journal of Oceanic Engineering 33(4) (2008), pp. 445–450. DOI: doi:10.1109/joe.2008.2001638.

[10] Davide Brizzolara Davide Anguita and Giancarlo Par- ODI. "Design and Implementation of HDL Modules and Circuits for Underwater Optical Wireless Communication". International Conference on Telecommunications and Informatics. Springer (9), 2009, p. 132.

[11] Aditi and Preeti. "Submarine Optical Cables as a Key Component in Undersea Telecommunications: A Review". In: International Journal of Application or Innovation in Engineering & Management (IJAIEM) 1.4 (Dec. 2012), pp. 79–83.

[12] Navneet Agrawal and Ajay Kumar Vyas. "Submarines Optical communication: A research review". In: International Journal of Electronics and Computer Science Engineering 1.2 (2012), pp. 370–374.

[13] Siang-Chih Chan et al. "Preliminary plan of underwater environmental monitoring in the offshore wind farm in the western sea of Taiwan". In: Challenges of Our Changing Global Environment. MTS/IEEE OCEANS -Bergen, 2013. DOI: doi:10.1109/oceans- bergen.2013. 6608093.

[14] I. Jurdana and V. Sucic. "Submarine optical networks: How to make them greener". In: 16th International Conference on Transparent Optical Networks (ICTON). 2014. DOI: doi:10.1109/icon.2014.6876663.

[15] Ammar A. Saleh, Amin B. A. Mustafa, and Ashraf A Osman. "Feasibility of Laying Fiber-Optic Cables underwater along River Nile Basin- Sudan Study Case". In: IOSR Journal of Computer Engineering (IOSR-JCE).17.1 (Jan. 2015), pp. 2278–8727.

[16] Erik Agrell et al. "Roadmap of optical communications". In: Journal of Optics 18.40 (June 2016), pp. 1– 40. DOI:/ 10. 1088 / 2040 - 8978/18/6/063002. URL:https://creativecommons.org/licenses/by/ 4.0/.

[17] D L Msongaleli et al. "Disaster-Aware Submarine Fiber- Optic Cable Deployment for Mesh Networks." In: Journal of Lightwave Technology 34(18) (2016), pp. 4293– 4303. DOI: doi:10.1109/jlt.2016.2587719.

[18] Abd-Elbaky, Mostafa, and Shuanggen Jin. "Estimating Runoff in the Nile River Basin from Multi-Satellite Measurements." In: 26th International Conference on Geoinformatics. 2018, pp. 1–5.

[19] Anna Tzanakaki et al. "Optical Network Design and Modeling". In: Optical Network Design and Modeling: 23rd IFIP WG 6.10 International Conference, ONDM 2019, Athens, Greece, May 13-16,

2019, Proceedings. Springer Nature, 2020.

[20] Tele-Geography. Submarine Cable Map. http:// www.subm arinecablemap .com / # /country / sudan. last accessed April 2020.

[21] Mark Van. DeRee and P Eng. “Submarine Cable Installation Techniques and Alternatives”, HEAVY MOV- ABLE STRUCTURES”. In: 14th Biennial Movable Bridge Symposium. October 22 – 25, 2012, ORLANDO, FLORIDA.

A Meta Analysis of Attention Models on Legal Judgment Prediction System

G.Sukanya[1], J.Priyadarshini[2]
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai Campus, Chennai, India

***Abstract*—Artificial Intelligence in legal research is transforming the legal area in manifold ways. Pendency of court cases is a long-lasting problem in the judiciary due to various reasons such as lack of judges, lack of technology in legal services and the legal loopholes. The judicial system has to be more competent and more reliable in providing justice on time. One of the major causes of pending cases is the lack of legal intelligence to assist the litigants. The study in this paper reviews the challenges faced by judgment prediction system due to lengthy case facts using deep learning model. The Legal Judgment prediction system can help lawyers, judges and civilians to predict the win or loss rate, punishment term and applicable law articles for new cases. Besides, the paper reviews current encoding and decoding architecture with attention mechanism of transformer model that can be used for Legal Judgment Prediction system. Natural Language Processing using deep learning is an exploring field and there is a need for research to evaluate the current state of the art at the intersection of good text processing and feature representation with a deep learning model. This paper aims to develop a systematic review of existing methods used in the legal judgment prediction system and about the Hierarchical Attention Neural network model in detail. This can also be used in other applications such as legal document classification, sentimental analysis, news classification, text translation, medical reports and so on.**

Keywords*—Legal judgment prediction; hierarchical attention neural network; text processing; transformer

I. INTRODUCTION

Legal Judgment Prediction (LJP) system helps in assisting litigants and attorneys to improve their work and time efficiency and reduce the risk of making mistakes with feasible judgment suggestions, which includes the prediction of charges, applicable law articles, and prison term based on the case facts [1]. Some of the LJP frameworks predict final judgment as a binary [2] and multilabel text classification [3] for cases in English of European Court Human Rights and Chinese Judgment Online dataset. One of the most important challenges in LJP is unlabelled data and that was tackled using the Long Short Term Memory framework [4] for Indian Supreme Court judgments with headnotes. At the initial stages, Machine Learning methods such as optimized Lasso Regression [5] [6] for Chinese cases and then deep learning models [7] [8] for automated judgment predictions were used. Providing fair and timely justice by the courts is not only the most important obligation of the country but is an important characteristic of democracy [9] [10] [11] [12] [13] [14]. India being the world's largest democracy is still in need of an intelligent judicial system. It also benefits civilians to know the possible judgment result before the trial by describing a case they are concerned about for win or loss rate. Surveys show that India has only twenty percent judges for every million citizens [15] [16].

Following are the benefits of using deep learning in LJP

- Accelerated decision process and outcomes.
- Instant verification of input data.
- Unbiased point of view or opinion.
- Ability to quickly showcase historical cases with similar patterns.
- Easier to spot corruption by identifying cases with high variance in human and AI decisions.

Case Facts of the real World has two main challenges. One is the difficulty faced in encoding lengthy documents and the other is lack of full external information. Existing models used for text classification and prediction like RNN and LSTM work sequentially and takes a long duration in training large corpus for the UK and Chinese court cases. Most of the LJP framework consider judgment prediction as text classification task while some works consider it as a Legal Reading Comprehension [3]. Any text classification has two main phases. The first phase is the representation of the document and the latter is to use a good classifier model. Both are very much important to give good prediction accuracy for new case facts or queries. Earlier deep learning models used BOW (Bag Of Words) and word embeddings, like Word2Vec, GloVe and Fast Text for encoding.

Legal information such as legal cases, contracts, bills are often represented in textual form. Processing of legal text is a grooming area in the current era. In NLP, the legal text is being utilized in various applications like "legal topic classification, court opinion generation and analysis, legal information extraction, legal interpretation and entity recognition" [17] [18] [19] [20] [21] [22]. Models that predict the legal outcomes are emerging for the past few years and these models aid the legal practitioners and citizens with a reduction in the cost of legal issues paves way for faster justice. The legal judgment prediction model can be utilized by lawyers and judges to predict the win or loss chance of a case [23] [24]. Human rights organizations and legal research scholars can adapt them to examine whether fair judicial decisions are given or are do they correlate with biases.

This review aims to discuss one of the most effective deep learning models (HAN) applicable to judgment predictions. The other objectives to be covered in this article are:

- Classify and summarize recent works based on empirical methods of LJP and conceptual literature of text classification.
- Encoding and Decoding architecture using a transformer.
- Showcase the important features and challenges of existing LJP methods.
- Future Scope with various applications.

The following section of this article is organized as follows. Section II briefs up the related work. In Section III, a review of the HAN and Transformer model is done. Major Findings and suggestions are revealed in Section IV. Finally, Section V contains the conclusion.

II. Background

Machine learning has begun to revolutionize various industries already and it would aid India's legal system in producing legal judgments with higher accuracy and precision. If law firms/advocates in India use this AI application for risk assessment for an out-of-court settlement/Alternative Dispute Resolution (ADR) mechanism, the number of cases would reduce and faster justice could be provided [25][26]. In India, legal search engines use Artificial Intelligence for legal analytics and visualization through its Case Map and Taxonomy and it also provides an analysis of judge's disposition through data analyzed by AI. Areas in legal that already use ML are client due diligence and contract management.

The problem of judgment prediction has taken a great attraction in the legal research area. This section describes Empirical Literature on Legal judgment Prediction methods, Conceptual Literature on Text Classification Methods and transformer model in detail.

A. Empirical Literature on Legal Judgment Prediction Methods

Most of the legal judgment prediction research works were classified using Binary Classification. Zhong *et al*. [1] have introduced the TOPJUDGE, a topological multi-task learning framework that shows the dependencies among subtasks of case facts as a Directed Acyclic Graph. The challenge faced in TOPJUDGE is that it failed to exhibit interaction between subtasks. Besides, a Convolutional Neural Network-based encoder was utilized to generate the fact descriptions. The outcomes of the TOPJUDGE model are law articles, charges and terms of penalty. TOPJUDGE exhibited a higher consistency over the existing models.

A new HAN model with Google's Bidirectional Encoder Representations from Transformers (BERT) was developed by Chalkidis *et al*. [2] for the prediction of cases from the European Court of Human Rights. A wide variety of neural models like BiGRU-Att, HAN, LWAN, BERT and Hierarchical BERT were evaluated on the proposed dataset. The research work had surpassed the drawbacks of the existing models and does (1) binary violation classification (2) multi-label classification (3) case importance prediction using sentence scores. Long *et al*. [3] formulated the Legal Reading Comprehension framework for the judgment predictions which works based on answering the questions of Comprehension rather than using the text classification method. This model was developed to handle multiple and complex textual inputs. Also, they have conceptualized the AutoJudge framework to infuse law articles for judgment prediction. The results of AutoJudge were better than the base model in terms of consistency and reliability. The problem of unavailability of labeled data for the prediction task is tackled by assigning classes/scores to sentences in the training set, based on their match with reference summary produced by humans using LSTM by Anand et al. [4].

In 2020, Guo *et al*. [5] introduced the TenLa by blending the concepts of both the tensor decomposition and an optimized Lasso regression model. Based on the similarities between legal cases the judgment charges were predicted. The major process undergone in the proposed works was: (a) ModTen to legal cases as three-dimensional tensors, (b) ConTen to decompose tensors obtained by ModTen (c) OLass, which was trained with the Core tensors got by ConTen. The results of the TenLa had exhibited higher accuracy than the traditional models.

Guo *et al*. [6] have preferred TenRR that amalgamates the tensor decomposition and ridge regression for judgment prediction of legal cases, and the proposed model had enclosed three major contributions. In the initial contribution, RTenr was developed as a tensor representation method to express the legal cases as three-dimensional tensors. In the second contribution, the ITend was introduced to decompose the original tensors representing legal cases into core tensors. In the contribution, the ORidge was built to construct an optimized Lasso judgment prediction model for legal cases. The results of the proposed work had exhibited higher accuracy than traditional methods for judgment prediction.

Chen *et al*. [7] analyzed the case description and predicted the judgment employing the deep learning model. The outcome of the deep learning model was in the form of three aspects: penalty, accusation and legal provisions. They predicted the latter aspects based on the FastText and TextCNN method. The resultant of the proposed judicial decision-making model was more accurate and persuasive. Yang *et al*. [8] proposed a Multi-Perspective Bi-Feedback Network. It had a word-level attention mechanism based on the topology structure among subtasks. Also a multi-perspective forward prediction and backward verification framework were designed to make use of the dependencies among multiple subtasks effectively. The word collocations features of fact descriptions were integrated into the proposed work to distinguish cases with similar descriptions but different penalties. The resultant of the proposed work had achieved significant improvements in terms of prediction accuracy. Shang Li et al. [12] discussed a Multichannel Attentive Neural Network(MANN) framework which predicts applicable charges, punishment terms and articles for Chinese court cases based on case facts for single defendant person using two-tier hierarchical architecture. K.

Zhu et.al [22] proposed Sequential Generation Network using a nested hierarchical attention mechanism for multi-charge prediction with single case defendants.

Kongfan Zhu et al [34] proposed Transformer-Hierarchical-Attention-Multi-Extra (THME) Network to extract the semantics of external information of the fact for prediction of Legal judgment based on multiple classes.

Jerrold sho et al [35] proposed a comparative study on NLP methods against statistical models on 6227 novel Singapore Supreme Court Judgments for the topic model, word embedding, and language model.

Hui Wang et al [36] have designed the LJP framework based on FastText and TextCNN for multilabel text classification for accusation prediction.

B. Limitations

We see that most of the existing legal judgment prediction system is applicable for a single defendant person and the judgments are predicted only based on case facts. In the real world, along with case facts, other external information such as evidence and emotions play a vital role in judgment which is a drawback found, which indirectly affects the prediction accuracy of the judgment. Also the problem of unavailability of structured legal data prone to an imbalanced dataset, gives a biased prediction. The lengthy case fact also is found to be a major challenge. Practically, Casefact with legal opinion comes around 60 to 100 pages. So a great amount of time is spent on extracting important points from them to make it into around 200 words per document either manually or by using a text summarization tool for a basic RNN model to work on it. In Table I, the features and challenges of existing Judgment prediction works are enlisted**.**

TABLE I. FEATURES AND CHALLENGES OF EXISTING PREDICTION WORKS

Author [Citations]	**Methodology**	**Data Sets Used**	**Features**	**Challenges**
Zhong *et al*. [1]	TOPJUDGE	**CJO**, **PKU**, and **CAIL**. CJO has criminal cases published by the Chinese government from China Judgement Online. PKU contains criminal cases published by Peking University Law Online CAIL(Chinese AI and Law Challenge)	✓ integrates multiple subtasks and make judgment predictions through topological learning framework ✓ judgment predictions through topological framework ✓ neural encoder for fact representation and subtasks with DAG dependencies.	✓ Limited to work on single defendants and charges. ✓ Need to explore how to infuse temporal factor into LJP
Chalkidis *et al*. [2]	HAN model with BERT	English legal judgment prediction dataset	✓ binary violation classification of articles ✓ multi-label classification of charges ✓ case importance prediction using sentence scores	✓ few-shot learning is not taken into account ✓ need to break the problem of charge prediction into different subtasks
Long *et al*. [3]	AutoJudge	Chinese Referee Document Network	✓ captures the complex semantic interactions among facts, pleas, and laws based on legal reading comprehension framework ✓ Improved F1 score, accuracy and precision	❖ don't have access to groundtruth law articles ❖ increases the computational complexity ❖ reduces the accuracy and stability
Anand *et al*. [4]	neural network	Indian Supreme Court Judgments (1947 to 1993)	✓ tackles the problem of unavailability of labeled data ✓ Uses Feed Forward Neural Network and Long Short Term Memory for case text summarization	❖ Higher cost ❖ Higher computational complexity ❖ Need sentence simplification approaches for complex and long sentences
Guo *et al*. [5]	TenLa	3,000,000 legal cases in the past five years from multiple provinces and cities in China	✓ higher accuracy ✓ removes redundant, meaningless, and inaccurate information	❖ Need to prevent over fitting ❖ Complex
Guo *et al*. [6]	TenRR	Chinese Referee Document Network	✓ greatly reduce the dimension of original tensors ✓ Removal of the meaningless and inaccurate information in original tensors	❖ Need to improve the accuracy of predictions
Baogui Chen *et al*. [7]	FastText and TextCNN	CAIL 2018 data set	✓ more accurate and persuasive decision-making ✓ Better in accuracy, and recall rate	❖ Need to improve the accuracy of model prediction
Yang *et al*. [8]	Multi-Perspective based BiFeedback Network (MPBFN) and a Word Collocation Attention (WCA) mechanism	Chinese AI and Law challenge (CAIL2018)	✓ improves the overall performance ✓ improve the performance of multitasking	❖ Need to reduce the misjudgment of penalty prediction

C. Conceptual Literature on Text Classification Methods

Before applying text classification methods text preprocessing has to be done. Preprocessing of raw text data gives good results on classification

1) Based on text preprocessing: Jin Wang et al [37] in 2019 implemented a regional CNN with LSTM model which comprises of two parts: to predict the Valency Arousal ratings of texts on Stanford Sentiment Treebank 1 dataset. The local information within the sentences is observed using regional CNN and long-distance dependencies are extracted by using LSTM across sentences that can be considered in the prediction process. Hao Fei et al [38] in 2020 finds multiple emotions using text as a multilabel classification problem using variational autoencoder and capsule module, to extract rich features in a sentence. Latent Topic attention-based routing algorithm is used in capsule module for pertaining the task. Zhang et al [39] in 2019 proposed a coordinated CNN-LSTM attention model to capture meaningful emotional dependent information, where filters of different widths are used in between word representation and pooling unit to get semantic information. Hao Peng et al [40] in 2019 uses a graphical capsule neural network model to capture rich information through a routing mechanism. This type of neural network model is found to be better than LSTM-RNN while considering long-term dependency. Zhongqing Wang et al [41] proposed the Hierarchical Attention Model in 2020 using Linguistic Attention based on argument representation, dependency representation and sentiment representation to extract meaningful words.

Word segmentation and tokenization are the initial steps in Natural Language Processing. The raw content of case fact description has to be preprocessed according to the application we choose. Mingjie Ling et al [46] use ELMo(Embeddings from Language Models) word embedding Language Model to overcome polysemy phenomena in word representation. The work of researchers Matthew E Peters et al [47] has proved that ELMO word embeddings are good to avoid polysemy phenomena than Skip-gram models and other word embeddings, like word2vec and GloVe which were widely used earlier. The main advantage of ELMo is that they have different word vectors under different contexts for the same word.

D. Transformer Model

Transformers which are at their budding stage in the application can be replaced for other methods in encoding and decoding text representation. They are pre-trained word embedding models used for text summarization and translation. Original transformer models have a large number of parameters and are compute-intensive. Also, they can be used for fixed-length documents only. Some of the transformer models are G-BERT, BioBERT, M-Bert, Trans-Bert, Clinical BERT, etc.

Jacob Devlin et al [42] in 2019 proposed deep bi-directional transformers by smoothing the existing pre-trained BERT model by adding one additional output layer, which is simple and powerful compared to the existing RNN models. Zhenzhong Lan et al [43] in 2020 proposed A Lite BERT which is better than BERT in terms of less memory consumption of the model by using two parametric reduction techniques. Chi Sun et al [44] in 2020 worked on different types of fine-tuning like single task and multitask tuning of parameters of the BERT for text classification. The tuned hyperparameters were then applied on eight different datasets and analyses were done. Zihang Dai et al [45] in 2020 proposed an attention model using XLnet which could learn 80% more dependency than RNN and better than existing vanilla transformers. The drawback is that the transformer model is highly compute-intensive and that it has a little struggle in handling negative sentences [47].

III. HAN and Transformer Model

Recurrent Neural Networks in deep learning models were widely used in Natural language processing tasks. Though it can capture contextual information over long distances compared to CNN, it suffers from the Vanishing Gradient and Exploding Gradient problem. While passing the information down between hidden layers during backpropagation, larger derivatives increase exponentially and then explode eventually creating Exploding Gradient problem. Similarly for smaller derivatives, the gradient decreases and vanishes eventually creating the Vanishing Gradient problem. To solve this semantic bias CNN with the max-pooling stage is adopted to get the most important information from text. Again approaches using the basic CNN model cannot represent the text semantically due to fixed window size. Other solutions for vanishing and exploding gradient problems are reducing the number of layers, by limiting the gradient size and by considering random initialization of weights [48] between the hidden layers. The gradient problems of RNNs were overcome by long short-term neural networks (LSTMs). It captures the contextual information of longer context in the documents than basic RNN. But, LSTM works unidirectional and sequentially which takes longer time consumption. This limitation in unidirectional LSTM was overriden by using bidirectional LSTMs, where we can read the context from both directions. Nowadays attention mechanisms were infused in the existing framework of RNN which is the hierarchical attention models. In this survey importance of the Hierarchical Attention Neural Network and transformer model has been studied and analyzed.

A. HAN

In NLP, the advancement in machine learning is making the decision-making capability a more relevant one. Bots in the market are already used to ‘smart search’ existing judgments and rulings to help in the preparation of a new case. Most of the existing judgment prediction models reviewed by researchers are based on traditional machine learning algorithms [28].

Nowadays, attention mechanism has been used in deep learning across a wide variety of contexts ranging from image captioning, image generation, and language modeling and translation. Hence, in the judgment prediction model, the attention mechanism is used to extract important words from the lengthy document, by assigning more weights to them.

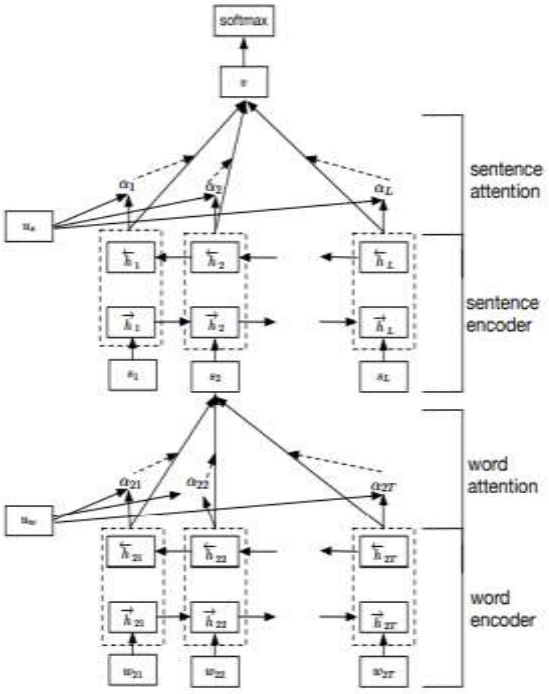


Fig. 1. General Architecture of Hierarchical Attention Network (HAN).

A hierarchical attention model is used in modeling the hierarchical relationships of words and sentences in a document for document classification. The Hierarchical Attention Network (HAN) [10] [22] [26] [27] [29] [30] [31] [32] [33] is a deep-neural-network that is utilized for Document Classification. A HAN attempts to classify a document based on the knowledge it can infer about the document from its composite parts, in other words, the sentences and words that make up the document. The 'hierarchical' in HAN comes from the design that this knowledge is built hierarchically, starting from using the words in a sentence and followed by using the sentences in a document [29].

The overall architecture of the Hierarchical Attention Network (HAN) is shown in Fig. 1. It consists of several parts: "a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level attention layer". The word vectors are encoded using the word sequence encoder and the word-level attention layer is utilized for aggregating the information of the informative words that do not contribute equally. Then, the sentence vectors are encoded in the word sequence encoder and the sentence level attention mechanism is utilized to reward attention (weights) to the sentences that are clues to correctly classify a document.

B. Transformer Model Architecture

Transformers have taken a vivid shape in NLP since 2019. Researchers use transformers nowadays in NLP while using RNN, LSTM, GRU, etc. Integration of transformer models with language models is at its budding stage in all NLP tasks. The Attention mechanism makes transformers have extremely long-term memory. A transformer model can remember all previously generated tokens that have been generated. Also, they have infinite reference windows, thereby overcoming the short reference window of RNN. It is an encoder-decoder architecture. The inputs given into the encoder are represented as a continuous vector.

The main benefit of using transformer-based models are:

1) The input tokens are not processed sequentially one by one as in RNN, rather the full sequence is taken as one input at a single shot.

2) Also labeled data is not necessary. Just giving a large amount of unlabeled data is enough to train a transformer-based model.

Bidirectional Encoder Representation Transformer (BERT) is a multiheaded attention-based encoder-decoder used as a pre-trained model for the word to a vector representation. It can be applied for Legal Judgment prediction which is based on lengthy case facts in an efficient way [2]. Fig. 2 shows the steps used in BERT architecture.

1) The input is fed into a word embedding layer where ever word is mapped into a vector and a lookup table is formed

2) Positional information is sent into embeddings since the encoder of the transformer doesn't have it. Sine and Cosine functions are used for positional encoding, at every even and odd index.

3) Encoder layer has the information for the entire sequence. It contains two subgroups. Multiheaded attention and then a Fully Connected Network. Multiheaded attention model uses a self-attention mechanism, i.e. it relates each word in input to other words. Query, key and Value factors are used to create a self-attention mechanism.

4) A dot product between Query and the key value is done to produce a score matrix. This gives a clue about how much importance should be given to each word in a sentence. Greater the score, the more important those words are. Thus queries are mapped to keys.

5) On the scaled score, a Softmax is applied which gives attention weights between 0 and 1.After doing this, greater scores are made still higher and vice-versa.

6) Multiply the above Softmax output with the value vector, which gives the output vector.

7) Decoder layer is also similar to the encoder layer and it has two multiheaded attention layers and a feedforward layer. It is appended with the linear layer that acts as a classifier and a Softmax is applied to get the probabilities of words. Masking is done to make all the negative values represented as zero.

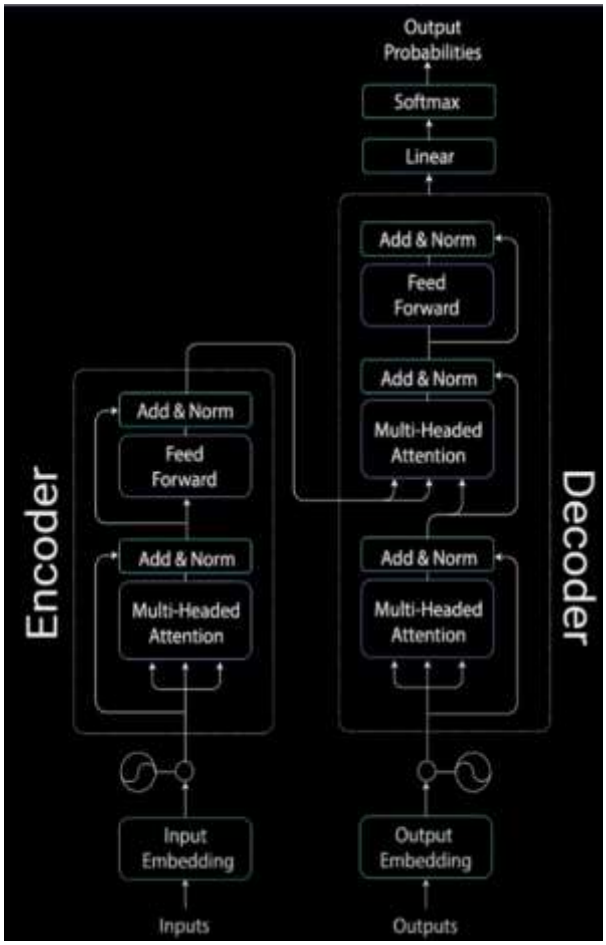


Fig. 2. Transformer Model.

IV. Major Findings

The explanation given in section III conveys the working procedure of HAN without transformer models and Transformer models with attention mechanism in a detailed way.

Hierarchical attention models are used mainly to automatically read and extract information from case facts efficiently. They differ from classic sequence to sequence model in two main ways. Most of the hierarchical model has two-tier architecture one for word attention vector and the other for sentence attention vector. The Encoder passes more data to the decoder instead of passing the last hidden state of the encoding stage, which turns to be advantageous than RNN. It consumes more time to train word embedding for a particular application if we use an attention model without transformers. Traditional RNN needs labeled data for the word which is not necessary if we use the pre-trained transformer model.

Attention mechanism with transformer model also has its pits and falls. The main benefits are it can cope up with any application using its good contextualized inbuilt word embedding which supports large words. A pre-trained transfer model in word to vector representation proves to be a time-saving one. On the other hand, original transformer(BERT) models are very big which are prone to intense computation. Due to the large number of parameters present in the transformer model, it is advised to fine-tune the architecture according to the needs of the application. BERT has its own limitations [26]. Firstly, it fails to capture longer-term dependency beyond the predefined context length. The maximum length of the sequence for BERT is 512 tokens which have to be taken into consideration. For shorter sequence padding has to be done and for longer sequence, the sentence has to be trimmed. Secondly, it struggles in handling negative sentences. Thirdly, it is unable to generalize to positions beyond those undergone for training.

There are different types of transformer models available. Though BERT has been renowned as the most efficient one on many NLP tasks, now it's overrun by XLNet from Google. XLNet uses the permutation language modeling concept in a sentence. CamemBERT is used mainly for legal tasks enduring with Part Of Speech tagging and Named Entity Recognition with less number of parameter compared to basic BERT, which in turn takes less time for computation. Pappagari et al. [26] proposed fine tuned BERT models such as Transformer over BERT(ToBERT) and Recurrence over BERT(RoBERT) methods for classification of long documents which performed better than pre-trained BERT. Therefore we suggest using HAN with fine-tuned transformer model for future endeavors in judgment predictions.

V. Conclusions

The purpose of this review was to identify an effective deep learning model used for judgment predictions. Based on the analysis conveyed integration of Hierarchical Attention Neural network models with fine-tuned transformer concept will give an efficient improvement based on quality and time in judgment prediction. Also, the improvement of multilabel classification for complex case facts with multiple defendants and charges still needs further investigation. A future exploration into the following legal areas such as summarization of legal judgment, legal data curation, and legal document simplification could be very much useful for the legal society.

References

[1] Haoxi Zhong, Zhipeng Guo∗ , Cunchao Tu, Chaojun Xiao, Zhiyuan Liuy , Maosong Sun,"Legal Judgment Prediction via Topological Learning",Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,pp.3540–3549,2018.

[2] Ilias Chalkidis, Ilias Chalkidi and Nikolaos Aletras, "Neural Legal Judgment Prediction in English",arXiv:1906.02059v1 [cs.CL],Jun 2019.

[3] Shangbang Long, Cunchao Tu, Zhiyuan Liu, Maosong Sun,"Automatic Judgment Prediction via Legal Reading Comprehension",arXiv:1809.06537v1 [cs.AI],Sep 2018.

[4] Deepa Anand, Rupali Wagh,"Effective Deep Learning Approaches for Summarization of Legal Texts",Journal of King Saud University - Computer and Information Sciences,2019.

[5] Xiaoding Guo,Hongli Zhang, Lin Ye,Shang Li,"TenLa: an approach based on controllable tensor decomposition and optimized lasso regression for judgement prediction of legal cases",Applied Intelligence,2020.

[6] X. Guo, H. Zhang, L. Ye, S. Li and G. Zhang, "TenRR: An Approach Based on Innovative Tensor Decomposition and Optimized Ridge Regression for Judgment Prediction of Legal Cases," in IEEE Access, vol. 8, pp. 167914-167929, 2020, doi: 10.1109/ACCESS.2020.2999522.

[7] Baogui Chen,Yu Li,Shu Zhang, Hao Lian,"A Deep Learning Method for Judicial Decision Support",IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C),2019.

[8] Wenmian Yang, Weijia Jia, Xiaojie Zhou and Yutao Luo,"Legal Judgment Prediction via Multi-Perspective Bi-Feedback Network",Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19),2019.

[9] D. Huang and W. Lin, "A Model for Legal Judgment Prediction Based on Multi-model Fusion," 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, pp. 892-895, 2019.

[10] S. Li, B. Liu, L. Ye, H. Zhang and B. Fang, "Element-Aware Legal Judgment Prediction for Criminal Cases with Confusing Charges," 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, , pp. 660-667, 2019.

[11] C. Wang and X. Jin, "Study on the Multi-Task Model for Legal Judgment Prediction," 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, pp. 309-313, 2020.

[12] S. Li, H. Zhang, L. Ye, X. Guo and B. Fang, "MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction," in IEEE Access, vol. 7, pp. 151144-151155, 2019. doi: 10.1109/ACCESS.2019.2945771.

[13] L. Chen, N. Xu and Y. Wang, "Legal Judgment Prediction with Label Dependencies," 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Calgary, AB, Canada,

[14] L. Yuan et al., "Automatic Legal Judgment Prediction via Large Amounts of Criminal Cases," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, pp. 2087-2091.pp. 361-365, 2020.

[15] R. Sil and A. Roy, "A Novel Approach on Argument based Legal Prediction Model using Machine Learning," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp. 487-490, 2020.

[16] X. Yang, G. Shi, J. Lou, S. Wang and Z. Guo, "Interpretable Charge Prediction with Multi-Perspective Jointly Learning Model," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, pp. 1850-1855, 2019.

[17] V. G. Pillai and L. R. Chandran, "Verdict Prediction for Indian Courts Using Bag of Words and Convolutional Neural Network," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 676-683, 2020.

[18] C. Jin, G. Zhang, M. Wu, S. Zhou and T. Fu, "Textual content prediction via fuzzy attention neural network model without predefined knowledge," in China Communications, vol. 17, no. 6, pp. 211-222, June 2020.

[19] T. Goto, K. Sano and S. Tojo, "Modeling Predictability of Agent in Legal Cases," 2016 IEEE International Conference on Agents (ICA), Matsue, pp. 13-18, 2016.

[20] Y. Yin, F. Zulkernine and S. Dahan, "Determining Worker Type from Legal Text Data using Machine Learning," 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Calgary, AB, Canada, pp. 444-450, 2020.

[21] K. Kowsrihawat, P. Vateekul and P. Boonkwan, "Predicting Judicial Decisions of Criminal Cases from Thai Supreme Court Using Bi-directional GRU with Attention Mechanism," 2018 5th Asian Conference on Defense Technology (ACDT), Hanoi, pp. 50-55, 2018.

[22] K. Zhu, B. Ma, T. Huang, Z. Li, H. Ma and Y. Li, "Sequence Generation Network Based on Hierarchical Attention for Multi-Charge Prediction," in IEEE Access, vol. 8, pp. 109315-109324, 2020.

[23] B. Chen, Y. Li, S. Zhang, H. Lian and T. He, "A Deep Learning Method for Judicial Decision Support," 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Sofia, Bulgaria, , pp. 145-149, 2019.

[24] J. Guo, B. Wu and P. Zhou, "BLHNN: A Novel Charge Prediction Model Based on Bi-Attention LSTM-CNN Hybrid Neural Network," 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Hong Kong, Hong Kong , pp. 246-252, 2020.

[25] Chang Yin,Cuiqing Jiang,Zhao Wang,"Evaluating the credit risk of SMEs using legal judgments",Decision Support Systems,2020.

[26] Raghavendra Pappagari, Piotr ˙Zelasko, Jes´us Villalba, Yishay Carmiel, and Najim Dehak,"Hierarchical Transformers For Long Document Classification", arXiv:1910.10781v1 [cs.CL] 23 Oct 2019.

[27] Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to Predict Charges for Criminal Cases with Legal Basis. arXiv preprint arXiv:1707.09168.

[28] Rafe Athar Shaikh,Tirath Prasad Sahu,Veena Anand,"Predicting Outcomes of Legal Cases based on Legal Factors using Classifiers",Procedia Computer Science,2020.

[29] D. Huang and W. Lin, "A Model for Legal Judgment Prediction Based on Multi-model Fusion," 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, pp. 892-895,2019.

[30] Sajad Mousavi,Fatemeh Afghah,U. Rajendra Acharya,"HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks",Computers in Biology and Medicine,2020.

[31] Fa Li,Zhipeng Gui,Yichen Lei,"A hierarchical temporal attention-based LSTM encoder-decoder model for individual mobility prediction",Neurocomputing, 2020.

[32] Yirong Zhou,Jun Li,Luo Chen,"A spatiotemporal hierarchical attention mechanism-based model for multi-step station-level crowd flow prediction",Information Sciences,2021-33.

[33] Shuning Xing,Fang'ai Liu,Tianlai Li,"A hierarchical attention model for rating prediction by leveraging user and product reviews",Neurocomputing,2019 -34.

[34] Kongfan Zhu, Rundong Guo , Weifeng Hu, Zeqiang Li, and Yujun Li "Legal Judgment Prediction Based on Multiclass Information Fusion" ,in Hindawi Complexity Volume 2020, Article ID 3089189, 12 pages,https://doi.org/10.1155/2020/3089189.

[35] Jerrold sho, Legal Area Classification: "A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments" Domains @agc.gov.sg @smu.ac.in.

[36] Hui Wang, Tieke He, Zhipeng Zou, Siyuan Shen, Yu Li "Using Case Facts to predict accusation based on deep learning",in 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C).

[37] Jin Wang, Liang-Chih Yu, Member, IEEE, K. Robert Lai, and Xuejie Zhang 11 December 2019, "Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis", Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28) pp: 581 – 591,Date of Publication: 11 December 2019 doi: 10.1109/TASLP.2019.2959251,Publisher: IEEE.

[38] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji,"Topic-Enhanced Capsule Network for Multi-Label Emotion Classification", Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28) Page(s): 1839 – 1848 Date of Publication: 10 June 2020 doi: 10.1109/TASLP.2020.3001390 Publisher: IEEE.

[39] Zhang Yangsen, Zheng Jia1 , Jiang Yuru, Huang Gaijuan And Chen Ruoyu ,"A Text Sentiment Classification Modeling Method Based on Coordinated CNN-LSTM-Attention Model" Published in: Chinese Journal of Electronics (Volume: 28 , Issue: 1 , 1 2019)Page(s): 120 – 126,Date of Publication: 22 August 2019,doi: 10.1049/cje.2018.11.004,Publisher: IET.

[40] Hao Peng, Senzhang Wang,Lihon Wong,Qiron Gong,"Hierarchical Taxonomy-Aware and Attentional Graph Capsule RCNNs for Large-Scale Multi-Label Text Classification", Published in: IEEE Transactions on Knowledge and Data Engineering (Early Access)Page(s):1–1, 2019,doi: 10.1109/TKDE.2019.2959991,Publisher: IEEE.

[41] Zhongqing Wang , Qingying Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou, "Neural Stance Detection With Hierarchical Linguistic Representations", Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28)Page(s):635–645,Date of Publication: 03 January 2020,doi: 10.1109/TASLP.2020.2963954,Publisher: IEEE.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Anthology ID:N19-1423 Volume:Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),June 2019, Minneapolis, Minnesota, Publisher:Association for Computational Linguistics, doi:10.18653/v1/N19-1423.

[43] Zhenzhong Lan, Mingda Chen, Sebastian Goodman1 Kevin Gimpel, Piyush Sharma Radu Soricut,"ALBERT: A Lite Bert For Self-Supervised Learning Of Language Representations" in arXiv:1909.11942v6 [cs.CL] 9 Feb 2020.

[44] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang, "How to fine-tune BERT for text classification?" , in arXiv:1905.05583v3 [cs.CL] 5 Feb 2020.

[45] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le2, Ruslan Salakhutdinov Carnegie Mellon University,GoogleBrain "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Count", arXiv:1901.02860v3 [cs.LG]2 Jun 2019.

[46] Mingjie Ling; Qiaohong Chen; Qi Sun; Yubo Jia,"Hybrid Neural Network for Sina Weibo Sentiment Analysis", Published in: IEEE Transactions on Computational Social Systems (Volume: 7, Issue: 4, Aug. 2020),Page(s): 983 – 990,doi: 10.1109/TCSS.2020.2998092 .

[47] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep contextualized word representations" in arXiv,preprint arXiv:1802.05365.

[48] J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity (2020), ICLR2020.

Development of a Virtual Pet Simulator for Pain and Stress Distraction for Pediatric Patients using Intelligent Techniques

Angie Solis-Vargas[1], Iam Contreras-Alcázar[2]
Escuela Profesional de Ingeniería de Sistemas
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

Jose Sulla-Torres[3]
Escuela Profesional de Ingeniería de Sistemas
Universidad Nacional de San Agustín
Arequipa, Perú

***Abstract*—Pediatric medical procedures are often stressful and painful for children, so they can resist and make the work of doctors and nurses a little more complicated. This research aims to develop a virtual pet simulator to distract pediatric patients from pain and stress using smart techniques. The methodology used is SUM. The primary data for the development of the simulator were gravity, the player's position, the speed, and the mass for the calculation of the predictive physics in the toy to interact with the pet. As part of the intelligent techniques, the A-star algorithm was used for the pet to follow the user and the flocking algorithm to have a natural behavior of a group of animals and thus have a higher immersion level. Trials were conducted with pediatric patients where those who made use of the virtual pet simulator during the medical procedure felt less pain and stress than those who did not try the simulator. Therefore, it is highly recommended to use alternatives such as the one developed to reduce pain and stress in pediatric patients.**

***Keywords*—Virtual pet; pediatric patients; pain; stress; smart techniques; A-star algorithm; flocking algorithm**

I. Introduction

Pain is an emotional and unpleasant experience that the pediatrician often encounters in his daily activities with his patients. To treat pain and anxiety in the best way is to avoid them. Therefore, it is essential to try to avoid anxiety and stress that causes painful sensations [1].

Child distraction appears to be the most widely used behavior management technique during medical procedures [2]. Pet therapy can also be used to distract pediatric patients in stressful or painful situations. Studies show that children who interact with therapy pets, typically dogs, show increased coping skills and reduced anxiety levels. [3].

A recent technological advance that can be an attractive distraction is virtual reality [4], which uses virtual environments. Virtual reality immersion systems are also known to allow the user to interact with the computerized environment they are viewing and the feasibility of using these computerized environments to reduce anxiety and pain associated with an invasive medical procedure in children with Cancer [4].

In a virtual environment, different types of things or virtual objects can be included; one type of these are pets. Virtual pets are common in the domain of children's games, where children play with the virtual pet and often interact with it via a button or touch interface [5].

In Peru, virtual environments and therapies with pets are scarce concerning the distraction of pain and stress of pediatric patients during stressful or painful medical procedures. It is known that a virtual environment can be of different types or modalities, one of them being those that include virtual pets. So, will a virtual pet simulator be able to distract pediatric patients from pain and/or stress during medical procedures using a virtual environment?

This study seeks to help pediatric patients who need a distraction from both the pain and/or stress that they suffer during medical procedures. Consequently, the goal is to develop a virtual pet simulator to distract pediatric patients from pain and stress during medical procedures using a virtual environment and smart techniques, as these can help improve the pediatric patient experience and assist them. To avoid trauma from high pain medical procedures and even avoid scarring pediatric patients during simple medical procedures.

To do this, the SUM methodology for video games was used to develop quality games in defined times and costs. The methodology goes through five phases: Concept development phase, planning phase, development phase, beta phase, and closing phase. Tools like Unity and Blender were also used. Besides, intelligent techniques such as the A-star algorithm were applied so that the pet can follow the user and the flocking algorithm to have a natural behavior of secondary fauna, and thus the level of immersion is better.

The article has more sections which talk about the following: Section II talks about the related works that provided information for the development of the virtual pet simulator. Section III contains the materials and methods developed for the development of the virtual pet simulator. Then in Section IV, we have the results obtained when using the developed simulator. In Section V we have the conclusions obtained after analyzing the individual results, and finally, we have in Section VI the future works to be developed.

II. Related Work

This section reviews work related to pain management, non-drug therapy, pet therapy, virtual reality for pain in medical procedures, and virtual pets.

A. Pain Treatment

The best treatment of pain and anxiety will be to avoid them by promoting prevention, anticipating the pain produced by diseases or procedures. It is also essential to a void anxiety and stress caused by the painful sensation [1]. To do this, you have to know how to value pain. In pediatric patients, it is complex to assess pain since they have a limitation or impediment to express, transmit or specify their pain (location, intensity, characteristics), and it is even more so while the age of the pediatric patient is younger [6].

To assess pain, it is useful in the emergency department to consider the child's process, changes in physiological parameters (increased heart rate and respiratory rate, cold skin, increased sweating, vasoconstriction of the skin), and the scales pain assessment. Different scales try to objectify the intensity of pain according to the age of the child. For those over three years of age, subjective scales are used. Ages 3-6 years old: color scales or facial drawings. Ages 6-12: numerical, visual analog, or color scales [1].

There are types of pain which we will mention below:

- According to the intensity of the pain: There is mild pain, which usually will suffice an analgesic drug administered orally, and moderate pain that may be necessary drug combinations and use, in addition to an analgesic, an anti-inflammatory, or a minor opioid [1].
- According to the pain duration: There is acute pain, which is pain directly related to a temporary injury and usually lasts for a short period (<6weeks). Also, chronic pain persists for a period longer than six weeks, often for months or years [7].
- Pain from therapeutic, diagnostic procedures: Currently, multiple procedures require pseudo-analgesia techniques. It is necessary to assess in these cases the degree of pain and anxiety that is going to be induced, to anticipate it [1].

Just as there are types of pain, there are also forms of pain measurement such as scales, which are presented below [8]:

- Visual analog scale (VAS): Measures pain intensity through a markerless line with lower or higher endpoints for pain intensity. This type of scale is used in school-age children.
- Numerical Rating Scale (EN): This scale is a variation of EVA. Use numbers (0-10 or 0-100) to rate the pain. This type of scale can be used in children seven years and older.
- Face, legs, activity, crying, and comfort (RPALC): This scale evaluates distressing behaviors in five categories (face, legs, activity, crying, and comfort). It is used to measure acute pain in children two months to 7 years.

Like pain, stress can also be measured using scales. Among the main ones are:

- Visual analog scale (VAS): VAS is also used to measure anxiety and consists of a horizontal line of 10 cm with a connection of two points to each other, where 0 is equivalent to "without worry or anxiety" and 10 indicates "the worst worry or anxiety" with opposite facial expressions joined along the same line. The child is asked to mark the point that best represents the anxiety he feels [9].
- The facial image scale (FIS): It was developed to assess dental anxiety status in children. It consists of 5 faces ranging from very happy to very unhappy, which children can easily recognize [10].

B. Non-Pharmacological Treatments

It has been shown that the reduction of pain and distress can be alleviated and that some simple, non-pharmacological techniques can alleviate the fear they cause. For example, ice or vibration therapy can also help distract patients during painful procedures. Pediatric pain relief devices in the shape of animals or insects are visually appealing to children for sale. Placed on your skin, they provide a cooling or vibrating effect that helps numb the injection pain. Pet therapy can also be used to distract pediatric patients in stressful or painful situations [3].

Sampson and Renee [3] mention some examples of distraction techniques for pediatric patients according to their age such as the calming effect of wrapping to which babies react well, soap bubbles or hide-a-face play that works well for young children, cartoons for preschool children, audio visual distractions for older children or teenagers, or also doing use of pet therapy for distract pediatric patients from stressful or painful situations.

C. Pet Therapies

Animal-assisted therapy, also known as "pet therapy," is the general term that refers to both animal-assisted activities and animal-assisted therapies [11]. The dog is ideal because he is more dependent on the human being and comes to learn and obey [12]. Studies show that children who interact with therapy pets, usually dogs, show increased coping skills and reduced anxiety levels [3]. Besides, Guha [13] mentions that the calming effects of animals are especially valuable with children.

In order to be able to carry out the therapies with pets, Sampson and Renee [3] mention that the policies and procedures of the hospital that uses this type of therapy must be followed, in addition to the approval of the parents because any type of contraindication must be ruled out such as allergies or the possibility of bacterial contamination

D. Virtual Reality for Pain in Medical Procedures

Virtual reality is useful in providing relief from acute and procedural pain and can help provide a corrective psychological and physiological environment to facilitate rehabilitation for pediatric patients suffering from chronic pain. Furthermore, virtual reality therapies that incorporate body movement tracking allow for greater interactivity [7].

A large body of evidence supports the efficacy of immersive virtual reality in reducing pain, anxiety and stress among pediatric patients undergoing burn care or cancer treatments, as it provides a means of human / human-computer interaction, in which a human becomes an active participant in a virtual environment created through a head-mounted display. Besides, by using virtual reality, the user actively participates in a virtual environment since real time changes with the user's movements [14].

E. Virtual Pets

According to Lin, Faas and Brady [15] a virtual pet is a type of agent that can be realistic or also abstract that is found in video games or virtual reality environments, they also highlight that people can have virtual pets in addition to or instead of a real pet for company, amusement or for simple distraction.

Virtual pets are common in the domain of popular children's games, where children play with the virtual pet and often nurture it through a button or touch interface, although more advanced interfaces feature gesture and speech interaction such as the game Kinectimals (Microsoft Xbox 360) or EyePet (Sony Playstation 3) [5].

Studies have shown that using and interacting with virtual pets to prevent the treatment of different diseases works quite well with children [5]. The Mixed Reality Virtual Pets system to reduce childhood obesity developed by Johnsen et al. [5] turned out to be very reliable, and the study was a resounding success despite having minimal game content, compared to much more elaborate entertainment games, the virtual pet managed to motivate the treatment group of children who exercised significantly more than his peers in the control group.

All these reviews of the related works have served to establish the basis for the virtual pet simulator's development proposal for pediatric patients.

III. Materials and Methods

The SUM methodology has been chosen for its advantages in developing video games, which is divided into phases and goes hand in hand with a risk management document. To better understand the scope of this project using the SUM methodology, the following diagram was followed, which can be seen in Fig. 1. The diagram has 5 phases, which we will talk about and explain what was developed in each of them to achieve the final product, i.e. the virtual pet simulator. All phases are accompanied by risk management.

A. Phase 1: Concept

In this phase, the project concept's development was carried out; in other words, the vision, genre, classification, characteristics, history, and setting of the simulator. This game's vision is to provide a virtual pet that provides entertainment to pediatric patients, and its genre is a simulation.

Then the classification was continued: type E (Everyone), that is, for everyone since it is a simulator for children. The gameplay of this simulator is as follows: When the user starts the game, a virtual pet will be presented in the first instance; To be exact a dog, the user will be able to interact with it in almost the same way as with a real one, that is, they will be able to: feed it, pet it and play with it.

What features does the simulator have? The simulator has the following:

- Be attractive in the eyes of children.
- Be a visual and auditory distractor.
- Be interactive by using the pet.

The pet was chosen to be a dog because children have a better interaction with them, as already mentioned in Section II of the article. Once the simulator's characteristics were established, the story of the game or simulator was developed, which is simple. It is a parallel world away from the anguish where the user can have fun with a lovely pet, distracting him from his fears.

Finally, the simulator setting was designed, which has a semi-forest, since it has a good number of trees but not too many to block the sunlight or generate too many shadows that could scare the user (pediatric patient). Flowers, grass, and stones were also placed to give a more rustic atmosphere. Finally, some villages were also located to avoid a feeling of being asked or of loneliness.

B. Phase 2: Planning

In this section, the development team members' roles for the rest of the project were defined. The number of iterations performed was also determined, and the milestones that had to be met were specified. In the same way, the objectives to be achieved to complete the project were defined. Finally, the characteristics of the video game were specified.

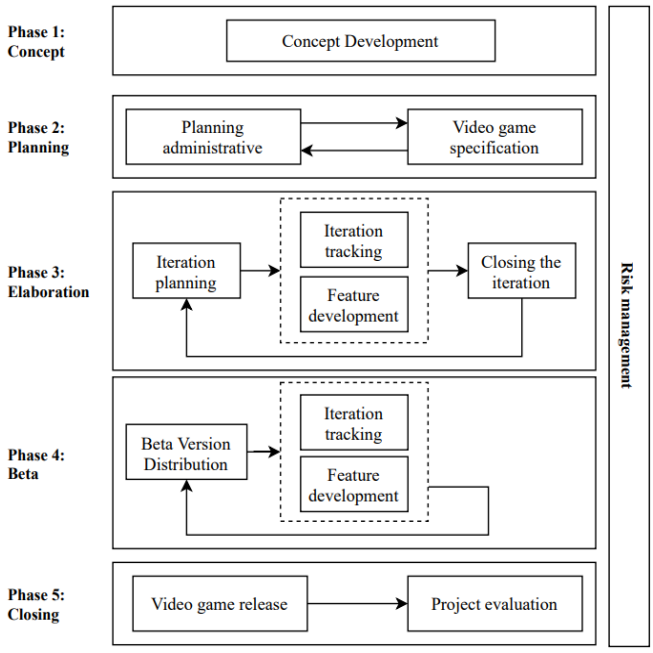


Fig. 1. SUM Methodology Diagram Designed by Acerenza et al. (2009) [16].

It was determined that the project would have three milestones which were developed in 17 weeks; therefore, a schedule was also developed. To better manage the project, the Trello tool was used to update the objectives or tasks that were concluded. As mentioned above, in this section, the objectives of the project were defined, which are:

- Generate a relaxed setting with a semi-forest environment.
- Generate a wide area for a good movement.
- Generate or obtain the model of a virtual dog.
- Generate functionalities with which you can interact with the virtual dog.
- Find and play calm and entertaining music for children.
- Insert sounds of interaction.

C. Phase 3: Elaboration

For the development of this project, it was divided into three stages. Stage 1 was the simulator analysis. For stage 2, the corresponding design was carried out, and finally, it culminates in stage 3, where the entire simulator implementation was carried out.

In the analysis stage, functional and non-functional requirements were defined. Below are some of the functionalities that were defined for the development of the project:

- The system will have the display function by using the first-person camera since the child must be able to see the pet because it is the main point of distraction.
- The system will have the interaction function when the child uses a controller to pet the pet.
- The system will have the functionality of displacement, which will be given through a command.
- The system will have the object manipulation functionality that will allow the child to interact with objects within reach in the area, such as balls, branches, or others to play with the pet.
- The system will have multimedia functionality since it will require sound or music to relax and distract the child more efficiently.

In the design stage, the project's sketches or mockups were made, which can be seen in the following figures. In Fig. 2, the game start tab is shown, which has the game title plus a button to start the game. Fig. 3 shows how the virtual dog is fed. In Fig. 4, it is shown that the user can take objects and throw them for the virtual dog to bring them back. In Fig. 5, it is shown that it is possible to interact with the virtual pet, that is, to caress it.

In this stage, the acquisition and/or creation of the necessary assets for the simulator implementation was also carried out. The asset of a 3D dog, quite friendly and attractive for children, was acquired from the Unity Asset Store as seen in Fig. 6, and the creation of assets of a toy bone was carried out as seen in Fig. 7 and a plate of food as seen in Fig. 8 using the Blender tool. It can be seen that the different models' figures are quite simple and with basic colors because children like simple shapes and attractive colors, so they were modeled based on said analysis.

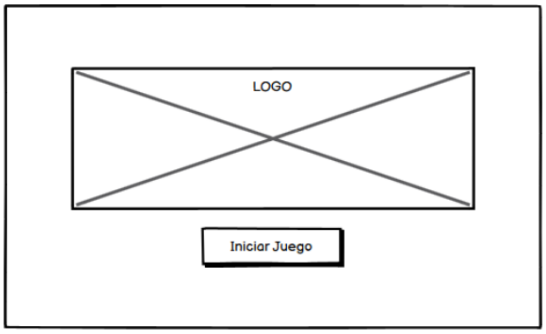


Fig. 2. Home Tab.

Fig. 3. Virtual Dog Feeding.

Fig. 4. Play with the Virtual Dog.

Fig. 5. Interaction with the Virtual Dog.

Fig. 6. Asset Virtual Dog.

Fig. 7. Asset Toy Bone.

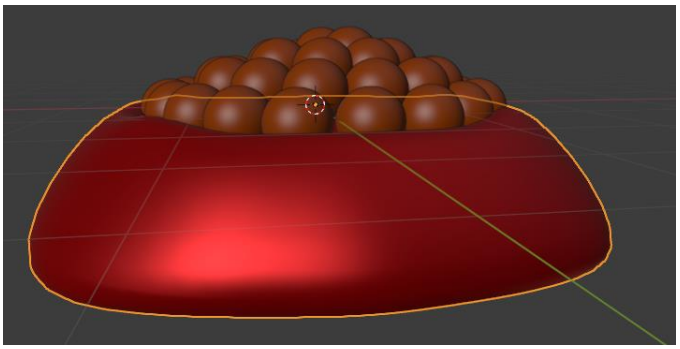


Fig. 8. Asset Food Plate.

In the implementation stage, the development of the scenario and different scripts in Unity was done. For the creation of the virtual terrain, the Gaia 2.0 tool was used; it helped speed up the development process a bit and generate a good quality scenario. The simulator scenario with that tool is presented below in Fig. 9. A forest was created in order to create a natural and relaxing environment because medical centers have unattractive rooms for children, even the fact of seeing a hospital, already begins to scare them, so the end of the virtual scenario of a forest, is to change the environment together with the background music, and thus generate tranquility.

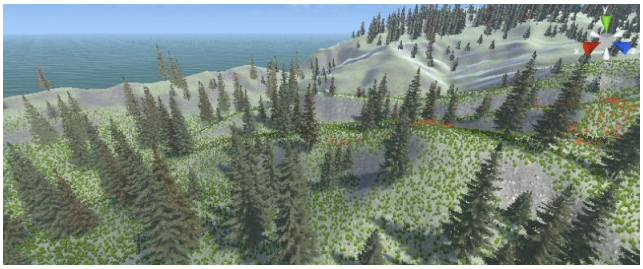


Fig. 9. Virtual Terrain using the Gaia Tool.

For the creation of scripts, the C # programming language was used since this language uses Unity. The launch script was developed, which allows the user to take a specific object from the stage and throw it; for this, gravity and the masses of the objects were taken into account to give them weight. The launch is based on the parabolic movement, and the prediction of the trajectory consists of the following elements: starting position in three dimensions based on time ($\bar{p}_t$); shooting position in three dimensions ($\bar{p}_0$); output speed (s_m); direction the toy was fired ($\bar{u}$); the length of time since the toy was released (t); and gravity ($\bar{g}$) which in this case has a value of 9.81 m/s. The corresponding formula can be seen in (1) with the active components.

$$\bar{p}_t = \bar{p}_0 + \bar{u}s_m t + \frac{\bar{g}t^2}{2} \quad (1)$$

In Fig. 10, you can see a cube, representing the toy (first part of development, the toy had not been modelled yet), being thrown openly on the stage; it is necessary to emphasize that the player can pick up this object, the longer he holds it, he will throw it; differently, it also depends on the mass of the rigid body component.

The launch script was made following the suggestion in Tuto_DrawTrajectory [17], where you can see the initial position of the toy with the applied force, through a loop "for" begins to go through and updates the points where the toy will go. Gravity is also applied based on the Unity game engine's physics component along with its time, as can be seen in Fig. 11.

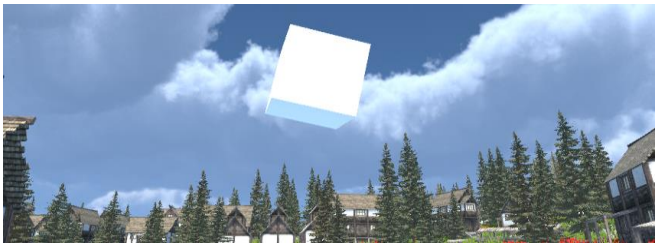


Fig. 10. Toy Launch.

```
public void UpdateDots (Vector3 toyPos, Vector2 forceApplied){
    timeStamp = dotSpacing;
    for (int i = 0; i < dotsNumber; i++) {
      pos.x = (toyPos.x + forceApplied.x * timeStamp);
      pos.y = (toyPos.y + forceApplied.y * timeStamp) -
        (Physics2D.gravity.magnitude * timeStamp * timeStamp) / 2f;
      dotsList [i].position = pos;
      timeStamp += dotSpacing;
    }
}
```

Fig. 11. Toy Launch Script.

For the smart part, the Pathfinding A-star algorithm was used, which is probably the most popular path search algorithm in artificial intelligence games [18]. The A-star algorithm was imported as a library to the virtual pet simulator in Unity. It is a generic search algorithm that can be used to find solutions for many problems, including route search. For route finding, the A-star algorithm repeatedly examines the most promising unexplored location it has seen. When a location is scanned, the algorithm terminates if that location is the target; otherwise, it takes note of all the neighbors at that location for further exploration.

The Pathfinding A star algorithm also made the virtual pet follow the bone toy, as shown in Fig. 12. The Pathfinding A-star library was implemented throughout the scenario; each node's size is one, since the problem was recognizing the terrain, specifically the houses, furthermore, it can be seen that the blue color is the space where the pet can move; near buildings there is no such color because it should not make sense for the pet to walk between the walls. The green line represents the shortest path of the grid drawn by the algorithm. The terrain is quite large because it has been created with Gaia 2.0, so the library's recognition dimensions had to be enlarged quite a bit and wait for it to recognize all the structures for the tour, which is shown in Fig. 13.

The Flocking algorithm was also used, to generate a more natural environment and with more significant nature, such as birds flying over a clear sky or a few butterflies flitting through the field, as can be seen in Fig. 14. The Flock and FlockUnit scripts were used for this algorithm. FlockUnit is in charge of giving individual flock behavior to a prefab assigned to it by the script; that is, it controls the cohesion, separation, and alignment components of a single prefab, and in case it has neighbors, that is, copies, it relates them to each other.

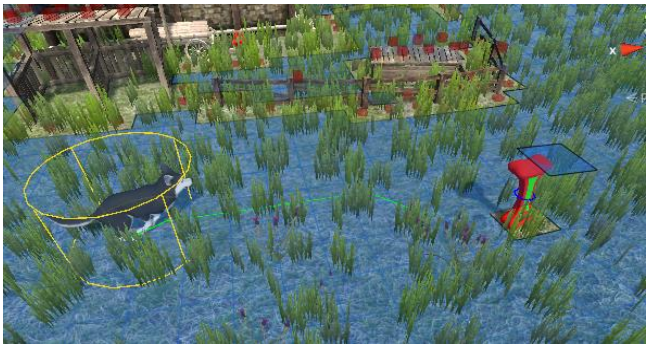


Fig. 12. Pathfinding Algorithm a Star.

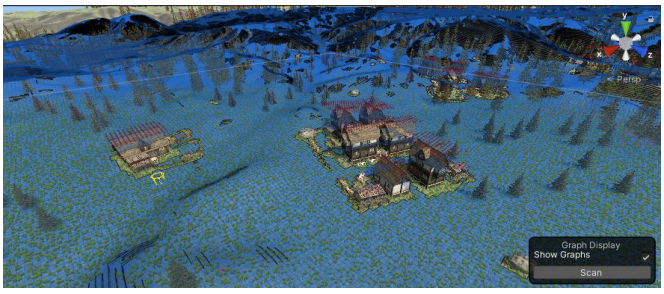


Fig. 13. Land with Pathfinding a Star Library.

Fig. 14. Scene using Flocking.

The Flock script is in charge of performing the flocking behavior of several copy objects that are asked to generate according to the type of prefab assigned using the relationship between neighbors of the FlockUnit script; that is, if it is assigned a bird prefab and it is instructed to generate 100 copies, then the Flock script will generate 100 bird copy prefabs, each of which will contain a behaviour given by the FlockUnit script that was assigned to the original prefab, and thus all the copies interact with each other as one giving the desired flock effect.

A life bar was also added using the Canvas tool, a tool already included within Unity. The life or energy bar represents the energy that the virtual pet has to be able to play with the user. If the life bar reaches the minimum, that is to say 0, and if the toy is out of reach, then the pet will lie down and rest until it is fed.

D. Phase 4: Beta

In the first version, it was obtained that the virtual pet pursues an object using the Pathfinding A star algorithm, the object that the virtual pet follows can be taken by the user and thrown. It was also possible to obtain a vast and natural scenery using the Gaia 2.0 tool since secondary nature such as flowers and trees were added. The flocking algorithm was also used to give it a more natural and immersive touch with secondary fauna such as birds in the sky and butterflies.

Two additional levels were generated for the better entertainment of pediatric patients if the user wishes to experience another type of entertainment. Level 2 is about collecting bones and achieving 100 points. The main level pet, which must be cared for and fed at this level, is manipulated by the user. The land is also in a natural style, including grass, stones, and trees. Different triggers were used to collect objects. Once the score is obtained, the level ends.

Finally, level 3 is focused on collecting clues to find the pet. The clues are three, which are a plate of food, a bone, and footprints. These clues are hidden on stage. This level's scenario is different from the main level and level 2 since it is a small city. Once the clues are found, the pet will appear, meeting with its owner (user).

E. Phase 5: Closing

In this phase, the problems, successes, solutions, achievement of objectives, and general feedback of the entire process of creating the video game were evaluated. To evaluate the aforementioned, verification tests were carried out to be more specific, unit tests.

Project verification tests were developed while adding a component to an item. The algorithms were evaluated according to their behavior based on the pet and the player. The respective unit tests were carried out on the entire project, but the most important or high-impact sections were: trajectory prediction, object tracking, petting the dog, and the flocking algorithm. Below you can see in Table I the white box tests performed to the trajectory prediction test section.

The project was also tested on two different computers. The first uses an AMD graphics card with a 6th generation Intel core i5 processor, while the other computer uses a GTX 1660 TI video card with an 8th generation Intel core i7 processor.

Metrics were carried out to know how much quality there is in the development of the project. The ISO \ IEC 9126 standard was used. From this standard, the characteristics of functionality, reliability, and usability were applied. For what are metrics, tables were made using the patterns required by the standard, that is, the tables contain the fields of name, purpose, application method, formula or measure, interpretation of the measured value, type of scale, type of measurement, source of measurement and audience.

For the functionality metrics, tests were carried out based on the requirements detected in the simulator development analysis stage. Table II shows the functionality metrics of one of the requirements.

In Table II, the interpretation of the measured value is 1, and for the functionality metrics, if said value is one or very close to it, it means that this requirement meets the functionality metrics. Similarly, the reliability and usability metrics were carried out, which can be verified in Table III, one of the reliability metrics of the trajectory prediction functionality, and Table IV, the usability metric of the interaction recognition functionality.

TABLE I. TRAJECTORY PREDICTION UNIT TEST TABLE

N°	Description	Input data	Expected output
1	User takes object	Left-click	Successful object take
2	The user moves with the object	Hold down the left click	Successful object take
3	User launches object	Release left click	Object dropped successfully
4	Object falls	Gravity, the mass of the object	Object drops successfully
5	Object collides with objects	Gravity, the mass of the object	The object collides and falls to the ground
6	Object collides with ceilings	Gravity, the mass of the object	The object stays on the ceiling
7	The object follows the launch path	Gravity, the mass of object, launch angle	The object follows the trajectory of the launch
8	The object is thrown with different force by pressing the mouse	Gravity, the mass of the object, force	Distance varies according to force
9	The object is thrown when the user jumps	Gravity, the mass of the object, user physics	The object can be thrown when the user jumps

TABLE II. TABLE OF FUNCTIONALITY METRICS OF THE FIRST REQUISITE

Name	Vision recognition functionality		
Purpose	How complete is the functional implementation of vision recognition?		
Application Method	I was counting the missing functions detected in the evaluation and comparing the number of functions described in the requirements specification performed in the analysis stage.		
Measure, formula and computer data	***"A" Number of missing functions***	***"B" Number of required functions***	***X***
	0	5	1
Interpretation of the measured value	1		
Scale type	absolute		
Media type	***A***	***B***	***X***
	numeric	numeric	1-(numeric / numeric)
Measure source	Specification of requirements in the analysis stage.		
ISO/IEC 12207 SLCP	Validation, Joint Review		
Audience	Analysts, developers		

TABLE III. TRAJECTORY PREDICTION RELIABILITY METRIC TABLE

Name	Trajectory prediction		
Purpose	How many of the required test cases are covered by the trajectory prediction test plan		
Application Method	Counting the missing functions detected in the evaluation of the trajectory prediction and compare with the number of functions described in the specification of requirements performed in the analysis stage.		
Measure, formula and computer data	***"A" Number of test cases in the plan***	***"B" Number of test cases required***	***X***
	8	9	0,8888888889
Interpretation of the measured value	0,8888888889		
Scale type	absolute		
Media type	***A***	***B***	***X***
	numeric	numeric	numeric/ numeric
Measure source	"A" comes from the test plan, "B" comes from the requirements specification		
ISO/IEC 12207 SLCP	Quality assurance, problem solving, verification		
Audience	Developers and maintainers		

It can be seen that Table II, Table III, and Table IV are close to or have one as an interpretation of the measured value. The functionalities or implementations complied with the metrics of ISO \ IEC 9126 [19]; therefore, there is quality in the virtual pet simulator developed.

TABLE IV. USABILITY METRIC TABLE OF INTERACTION RECOGNITION

Name	Interaction recognition functionality		
Purpose	What proportion of the system functions are evident to the user to have a good recognition of the interaction		
Application Method	Counting the functions evident to the user and compare with the total number of functions		
Measure, formula and computer data	***"A" Total of obvious functions***	***"B" Total number of functions***	***X***
	4	5	0,8
Interpretation of the measured value	0,8		
Scale type	absolute		
Media type	***A***	***B***	***X***
	numeric	numeric	numeric/ numeric
Measure source	Requirements specification, design		
ISO/IEC 12207 SLCP	Validation, Joint Review		
Audience	Analysts, developers		

Once the verification tests had been concluded and the simulator's quality had been verified, the validation tests were carried out. For these tests, a small medical center (medical post) was attended. A small number of 10 pediatric patients were evaluated, which we detail later, who have been prescribed an invasive medical procedure (injectable or serum). 4 children were taken as a control group. Six as a test group, it is known that most children feel pain and stress due to injectable, so four children were chosen as a control group since three were very few and what was wanted was to have an experimental group of as many of the small sample as possible to demonstrate the effectiveness of the pet simulator. Having a small sample is for reasons of the covid-19 of the year 2020 since this study was carried out during that year, in addition, the medical field in our country is collapsing, for which there is not much medical attention in external clinics and an enter to emergency admission is very risky for our health.

To verify the pet simulator's effectiveness, methods and scales were used to measure pain and stress in patients, which are: The Facial Image Scale (FIS) for measuring stress and the Visual Analog Scale (VAS) for both pain and stress.

Vital signs (heart rate, respiratory rate) of pediatric patients were also evaluated to have more objective results. These signs were noted and measured before and after the medical procedure for all pediatric patients. In this way, it was observed how much the heart rate and/or respiratory rate varied. Heart rate was measured by counting the pulses in the radial artery for 1 minute. Taking the vital signs of pediatric patients, the effectiveness of the virtual pet simulator could be verified.

The measurement scales were given to the pediatric patients through sheets to mark or indicate the way they felt during the medical procedure, both those who used the simulator or not. Below it can see in Fig.15 and Fig.16 the scales in the EVA, FIS figures, which were explained in Section II.

The tests were carried out in the medical center for six days, in which the informed consent of the head and the parents' respective permission with four boys and six girls were requested. Two of them are eight years old, one nine years old and the other ten years old. In the girls, four were five years old, one was four years old, and one was nine years old.

Each of the children attended the medical center and was diagnosed to receive invasive medical procedures. In the first four days, five pediatric patients were diagnosed with invasive medical procedures, of which the first four were taken as a control group, and from the 5th pediatric patient, they were taken as an experimental group.

The first four pediatric patients were three girls (two 5 years old and one 4 years old) and one boy (8 years old), who, as mentioned above, were selected as a control group; therefore, they did not use the simulator during the invasive medical procedures. The control group was evaluated for vital signs (pulse, respiratory rate) before and after the medical procedure, and they were given the respective pain and stress measurement scales (VAS, FIS).

The experimental group consisted of 6 pediatric patients, three girls (two 5-year-olds and one 9-year-old) and three 8, 9, and 10-year-old boys. This group did use the simulator during the invasive medical procedure by putting on virtual reality glasses type Vr Box 2nd Generation. Before performing the invasive medical procedure, when the pediatric patient had accelerated vital signs due to stress, they were given a simulator and expected to interact with the pet. Once the pediatric patient was distracted, the injectable was applied. Finally, once the invasive medical procedure was completed, the use of the simulator was withdrawn.

As was done in the control group, the pediatric patients in the experimental group were assessed for vital signs (heart rate, respiratory rate) both before and after medical procedures. Finally, after carrying out invasive medical procedures, the experimental group was given the respective pain and stress measurement scales (VAS, FIS).

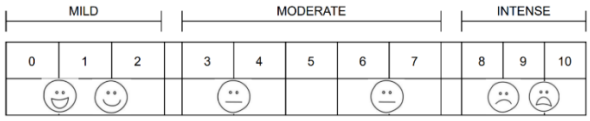


Fig. 15. Visual Analog Scale (VAS).

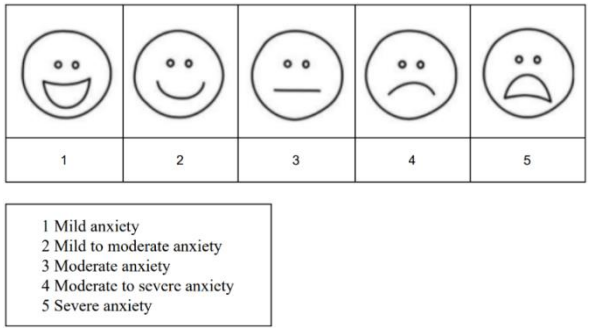


Fig. 16. Facial Image Scale (FIS).

IV. Results

A. Verification Test Results

After running the virtual pet simulator, it was observed that the algorithms work correctly; the trajectory prediction algorithm works at 89% since sometimes the object collided with the roof of a house and got stuck there, the pet followed it but could not climb and reach it, the toy must be set to recognize whether it has a considerable difference from the vertical coordinate of the pet, and make the toy fall.

In the movement algorithms, as for the pathfinding. A star algorithm, it was observed that the pet follows the character by 90% because there is a lake in the simulator, and if the user enters said lake, the pet should stop at wait for it, however, it follows it under the water, it can also modify the animation, in such a way if it detects that it is going to enter a water area, the pet maintains its vertical coordinate and proceeds to swim while following the player.

As for the flocking algorithm, the birds and butterflies were correctly added, these objects rise to the sky, and the simulator's immersion increases. As for its function the algorithm is at 100% because in the sky there, are no collisions with rocks or mountains, except for the same copy objects (neighbors); that is, the flocking algorithm works correctly.

B. Validation Test Results

First, the results obtained in the control group are presented. Pediatric patients' vital signs were measured before and after the medical procedure, as explained above. In Table V, you can see the vital signs of pediatric patients in the control group before and after performing the medical procedure.

Typical heart rate values in pediatric patients are detailed in Table VI [20].

The typical respiratory rate values in pediatric patients are detailed in Table VII [21].

After the medical procedure, it can be seen that in Table V that pediatric patients came to present an increase in both heart and respiratory rates. These results demonstrate that pediatric patients in the control group experienced both pain and stress while undergoing the invasive medical procedure. Pediatric patients were also given files or surveys containing the aforementioned scales.

For the VAS scale, pediatric patients marked values from 6 to 9, which means that pediatric patients felt pain intensely and moderate. For the FIS Scale, pediatric patients indicated faces from number 3 to 5, which means that all pediatric patients had moderate to even severe anxiety.

Now the results obtained in the experimental group are presented. The pediatric patients' vital signs in the experimental group were also measured before and after the medical procedure. In Table VIII, you can see the pediatric patients' vital signs in the experimental group before and after performing the medical procedure.

After the invasive medical procedure, Table VIII shows that the heart rate values and the respiratory rate decrease, compared to the values measured before applying the developed virtual reality environment. Therefore, the results showed that pediatric patients could distract themselves from the pain and stress of the situation surrounding them, which was the realization of the invasive medical procedure, giving more attention to the virtual pet simulator.

TABLE V. Pediatric Control Group Vital Signs Table

Nº	**Age**	**Sex**	**Heart frequency**		**Breathing frequency**	
			Before	***After***	***Before***	***After***
1	5	F	124	157	35	58
2	5	F	125	154	37	56
3	4	F	120	159	40	60
4	8	M	110	146	25	45

TABLE VI. Chart of Heart Rate in Pediatric Patients

Age	**Lower limit**		**Upper limit**	
2 age	80		130	
4 age	80		120	
6 age	75		115	
8 age	70		110	
10 age	70		110	
	Girls	***Boys***	***Girls***	***Boys***
12 age	70	65	110	105
14 age	65	60	105	100
16 age	60	55	100	95
18 age	55	50	95	90

TABLE VII. Table of Respiratory Rate in Pediatric Patients

Age	**Lower limit**	**Upper limit**
1–3 age	24	40
4–5 age	22	34
6–12 age	18	30
13–18 age	12	16

TABLE VIII. Table of Vital Signs of the Pediatric Experimental Group

Nº	**Age**	**Sex**	**Heart frequency**		**Breathing frequency**	
			Before	***After***	***Before***	***After***
1	8	M	112	90	29	18
2	5	F	129	112	37	28
3	10	M	107	70	25	17
4	5	F	131	115	37	30
5	9	F	109	93	30	20
6	9	M	106	87	26	19

V. Conclusions

It was found that the application of the virtual pet simulator allows distracting pediatric patients from pain and stress, according to the results obtained in the tests carried out. The tests showed that the pulse and respiratory rate levels of 60% of the total pediatric patients decreased since six pediatric patients made use of the pet simulator and all six managed to be distracted, that is to say, 100% of pediatric patients in the experimental group, they managed to distract themselves from pain and stress.

All the pulsations of pediatric patients before the medical procedure exceeded 100 beats per minute; also, all pediatric patients' respiratory rate exceeded 26 breaths per minute. However, after using the pet simulator, those of the experimental group, both their pulsation levels and their respiratory rates, decreased considerably.

It was also found that using the pet simulator would improve care during medical procedures for pediatric patients. Therefore, the virtual pet simulator developed allows the distraction of pain and stress for pediatric patients in a virtual environment; for this, the product was developed with the SUM methodology. The operation of the game has excellent playability since all the algorithms work as expected. Besides, thanks to intelligent techniques such as the Pathfinding A star algorithm, which allowed a better interaction between the pet and the user (pediatric patient), and the Flocking algorithm, which allowed a more natural environment, a better immersive environment was obtained. Finally, the simulator passed the metrics of ISO \ IEC 9126; therefore, the virtual pet simulator developed has quality.

VI. Discussions

As we can see, the heart rate, respiratory rate and anxiety in patients decrease. This is useful in pediatric patients, since post-traumatic stress can be avoided after a medical intervention that is very traumatic for the child, this has been the main motivation for this work. There are pathologies in which an increased heart rate can lead to a series of complications as in the case of congenital heart disease [22]; in addition, there are more complicated diseases such as the case of autistic children that are also affected by these variations [23]. This simulator can solve these complications by reducing heart frequency, respiratory frequency, and state of anxiety when these people are undergoing medical intervention.

We can observe in the work of Buldur and Candan [24] that they developed a software similar to ours in the field of virtual reality, however they work with other biological parameters while in ours, we add a new parameter, where we can observe a considerable decrease in respiratory rate.

Although modern systems use different technologies, in our environment the use of virtual reality is just emerging and in the different medical centers they are not yet used, they perform different traditional treatments. Therefore, the proposed simulator would help to encourage the use of modern technology in our health environments and thus obtain better results in the treatment.

VII. Future Work

The simulator is still in version 1.0. It can be improved since more details can be added, such as that the pet can receive orders such as sit or play dead. It is also intended to make an improvement to level 3 by placing the clues randomly since the present hidden clues are static so that if the child plays the level several times, he will be able to memorize the objects' location.

It is planned to test the virtual pet simulator with a larger sample when the COVID-19 situation improves to have better results and check its effectiveness. It is also planned to test the simulator with other medical procedures because it was only tested with invasive medical procedures.

Finally, it has been thought to generate more levels to the simulator; that is, it will not be oriented only for children, but also for different types of people, creating levels of distraction and stress for different ages since not only children suffer pain and stress during medical procedures, also young adults and older adults.

References

[1] J. Travería Casanova, T. Gili Bigatá and J. Rivera Luján, "Tratamiento del dolor agudo en el niño: analgesia y sedación", Servicio de Pediatría. Hospital de Sabadell, vol. 2, p. 22, 2010. [Accessed 18 October 2020].

[2] N. Asl, L. Erfanparast, A. Sohrabi, S. Ghertasi and A. Naghili, "The Impact of Virtual Reality Distraction on Pain and Anxiety during Dental Treatment in 4-6 Year-Old Children: a Randomized Controlled Clinical Trial", Dental Research, Dental Clinics, Dental Prospects, vol. 6, no. 4, p. 8, 2012. Available: 10.5681/joddd.2012.025 [Accessed 17 October 2020].

[3] J. Sampson and R. Allbright, "Distraer a los pacientes pediátricos durante los procedimientos dolorosos", Preguntas Clínicas, p. 2, 2019. [Accessed 18 October 2020].

[4] J. Gershon, E. Zimand, M. Pickering, B. Olasov and L. Hodges, "A Pilot and Feasibility Study of Virtual Reality as a Distraction for Children With Cancer", Journal of the American Academy of Child & Adolescent Psychiatry, vol. 43, no. 10, pp. 1243-1249, 2004. Available: 10.1097/01.chi.0000135621.23145.05 [Accessed 17 October 2020].

[5] K. Johnsen *et al.*, "Mixed Reality Virtual Pets to Reduce Childhood Obesity", IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 4, pp. 523-530, 2014. Available: 10.1109/tvcg.2014.33 [Accessed 19 October 2020].

[6] M. Narváez Tamayo, "Treatment of pain in childrens", Educación Médica Continua, vol. 49, no. 1, pp. 66-74, 2010. [Accessed 18 January 2021].

[7] A. Stevenson Won, J. Bailey, J. Bailenson, C. Tataru, I. Yoon and B. Golianu, "Immersive Virtual Reality for Pediatric Pain", children, vol. 4, no. 52, p. 15, 2017. Available: 10.3390 [Accessed 18 October 2020].

[8] A. Bayat, R. Ramaiah and S. Bhananker, "Analgesia and sedation for children undergoing burn wound care", Expert Review of Neurotherapeutics, vol. 10, no. 11, pp. 1747-1759, 2010. Available: 10.1586/ern.10.158 [Accessed 23 January 2021].

[9] K. Wolitzky, R. Fivush, E. Zimand, L. Hodges and B. Rothbaum, "Effectiveness of virtual reality distraction during a painful medical procedure in pediatric oncology patients", Psychology & Health, vol. 20, no. 6, pp. 817-824, 2005. Available: 10.1080/14768320500143339 [Accessed 18 October 2020].

[10] Y. Matsuoka and K. Fukai, "Face Scales and Facial Expression Analysis to Assess Clinical Pain Intensity", vol. 8, no. 1, p. 8, 2008. [Accessed 23 January 2021].

[11] A. Tielsch Goddard and M. Jo Gilmer, "The Role and Impact of Animals with Pediatric Patients", Pediatric Nursing, vol. 41, no. 2, p. 7, 2015. [Accessed 19 October 2020].

[12] R. Zapata, E. Soriano, A. González, V. Márquez and M. López, "Influencia de las mascotas en los niños", in Educación y salud en una sociedad globalizada, R. Zapata, E. Soriano, A. González, V. Márquez and M. López, Ed. Universidad de Almería, 2015, p. 6.

[13] M. Guha, Handbook on animal-assisted therapy: Theoretical foundations and guidelines for practice, 2nd ed. London: JOURNAL OF MENTAL HEALTH, 2012, pp. 320-323.

[14] C. Lee, M. Wa and K. Chow, "Effects of immersive virtual reality intervention on pain and anxiety among pediatric patients undergoing venipuncture: a study protocol for a randomized controlled trial", Trials, vol. 20, no. 1, 2019. Available: 10.1186/s13063-019-3443-z [Accessed 17 January 2021].

[15] C. Lin, T. Faas and E. Brady, "Exploring affection-oriented virtual pet game design strategies in VR attachment, motivations and expectations of users of pet games", 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), 2017. Available: 10.1109/acii.2017.8273625 [Accessed 17 October 2020].

[16] N. Acerenza *et al.*, 2009. Una Metodología para el Desarrollo de Videojuegos. Simposio Argentino de Ingeniería de Software, pp.171-176.

[17] "Tuto_DrawTrajectory", GitHub, 2020. [Online]. Available: https://github.com/herbou/Tuto_DrawTrajectory/blob/master/Assets/Scripts/Trajectory.cs. [Accessed: 19- Oct- 2020].

[18] X. Cui and H. hi, "A*-based Pathfinding in Modern Computer Games", IJCSNS International Journal of Computer Science and Network Security, vol. 11, no. 1, p. 6, 2011. Available: https://www.researchgate.net/profile/Xiao_Cui7/publication/267809499_A-based_Pathfinding_in_Modern_Computer_Games/links/54fd73740cf270426d125adc.pdf. [Accessed 17 November 2020].

[19] Ingeniería de software—Calidad del Producto— Parte 1: modelo de la calidad (ISO/IEC 9126-1:2001, IDT), 1st ed. Habana, 2005, p. 33.

[20] R. Kliegman *et al.*, Nelson. Tratado de pediatría, 21st ed. Barcelona: Elsevier, 2020.

[21] D. Cobo and P. Daza, "Signos vitales en pediatría", Revista Gastrohnup, vol. 13, no. 1, 2011. [Accessed 16 January 2021].

[22] P. Yubbu, N. Devaraj, D. Sahadan and H. Latiff, "Vascular compression of the airways: Issues on management in children with congenital heart disease", Progress in Pediatric Cardiology, vol. 59, p. 101207, 2020. Available: 10.1016/j.ppedcard.2020.101207 [Accessed 24 February 2021].

[23] R. Thapa et al., "Heart Rate Variability in Children With Autism Spectrum Disorder and Associations With Medication and Symptom Severity", Autism Research, vol. 14, no. 1, pp. 75-85, 2020. Available: 10.1002/aur.2437 [Accessed 24 February 2021].

[24] B. Buldur and M. Candan, "Does Virtual Reality Affect Children's Dental Anxiety, Pain, And Behaviour? A Randomised, Placebo-Controlled, Cross-Over Trial", Pesquisa Brasileira em Odontopediatria e Clínica Integrada, vol. 21, 2021. Available: 10.1590/pboci.2021.002 [Accessed 24 February 2021].

Identifying Communication Issues Contributing to the Formation of Chaotic Situation: An AGSD View

Hina Noor[1], Babur Hayat Malik[2], Zeenat Amjad[3], Mahek Hanif[4]
Sehrish Tabussum[5], Rahat Mansha[6], Kinza Mubasher[7]
Department of CS and IT, University of Lahore, Gujrat Campus, Pakistan

***Abstract*—The software can be constructed in many different contexts using various approaches to software creation, Software Development (GSD), Agile Software Development (ASD) and Agile Global Software Development (AGSD) in an ecumenically distributed way (a coalescence of GSD and ASD). This GSD (Global Engenderment of Software) is becoming increasingly important. Although communication is important in the sharing of information between team members, there are additional barriers to multi-site software creation, various time zones and cultures, IT infrastructure, etc., and delays in communication activities that are already problematic. In the case of Agile Global Software Development (AGSD), Agile Global Software Development (AGSD) is much more critical and plays a primary role in interaction and communication. The aim of this paper is to tackle the chaos problems associated with evolution of Agile Global Software (AGSD). We have obtained knowledge from previous works and from web reviews from worldwide, a literature review was conducted. Using a conceptual model, tabulated based on authors, and addressed also, the chaos issues are then illustrated. We identify the most discussed and less discussed issues in the literature. It is consequential to define the chaos issue in order to illustrate the genuine issues that subsisted in AGSD.**

Keywords*—*Chaotic situation; chaos; issues; communication; agile; distributed software development; global distributed software development; communication challenges; AGSD

I. INTRODUCTION

The aperture among time and location between dispersed software advancement groups is right now considered to integrate to the clamorous Ecumenical Software Development environment (GSD). In addition, as it includes using the agile process, coordination between developers and customers is essential [1]. "Development of software with teams located at various geographical locations, from different national and organizational cultures and time zones" Distributed Software Development (DSD), Ecumenical Software Development is kenned for this kind of development (GSD) [24], [37]. Ecumenical Distributed Software Development or Ecumenical Software Development (GSD) has now become the standard for the Ecumenical market in the software industry. This phenomenon is the product of global economic globalization, which provides the tech industries around the world with a forum for global competitive advantage. In producing quality tech, software companies are competing with each other. With the presence of IT, software creation anywhere in the world, it can be done. Distributed development teams of diverse backgrounds, cultures and languages can be based in different time zones in different geographical areas and still function on the same project [12]. Global team issues shown in Fig. 1.

Over the earlier decade, worldwide programming advancement (GSD), IT rethinking, and reevaluating of business measures have demonstrated yearly development paces of 10-20 percent [5], [6]. The level of offshoring or globalization relies upon the details of the hidden business and what programming is being created. For example, albeit the dispersed IT application made is genuinely circulated, the advancement of straightforward installed programming frameworks actually faces huge difficulties when executing appropriated amplification [13], [14].

The implementation of limber software development has brought paramount amendments to the world of software engineering over the last decay. Many companies and software developers are now utilizing nimble strategies to engender the most efficacious and in the shortest time possible, high-quality product. Extraordinary programming (XP), scrum, lean programming made, include driven amplification (FDD), and DSDM and precious stone approaches are the various kinds of deft procedures utilized by these associations [24]. For the situation of deft ecumenical programming advancement (AGSD), where correspondence assumes an essential part, communication is even more paramount. According to the Supple Manifesto, “Throughout the project, business people and developers must collaborate circadian”. In AGSD, communication quandaries were withal widely addressed in the literature [2], [15], which showed that distributed teams, categorically supple teams; rely heavily on communication implements [4]. Albeit several studies have explored the optimal technical stool for fortifying efficient communication in AGSD, it is still an unresolved quandary [3], [16]. As, agile development is elaborated in Fig. 2.

In contrast to geographically dispersed GSD teams, limber development fixates on active face-to-face contact between co-located teams. Limberness provides GSD with both advantages and challenges, such as the difficulty of communication. As interest in using nimble GSD techniques has increased, there has also been an increase in communication literature, as well as communication techniques and limber GSD strategies. There is a desideratum in versatile GSD for researching communication dilemmas and creating or using tools, strategies, and methods to tackle them [7]. By defining, synthesizing and presenting the communication dilemmas of versatile GSD, the purpose of this research paper is to address the above-mentioned gap [17].

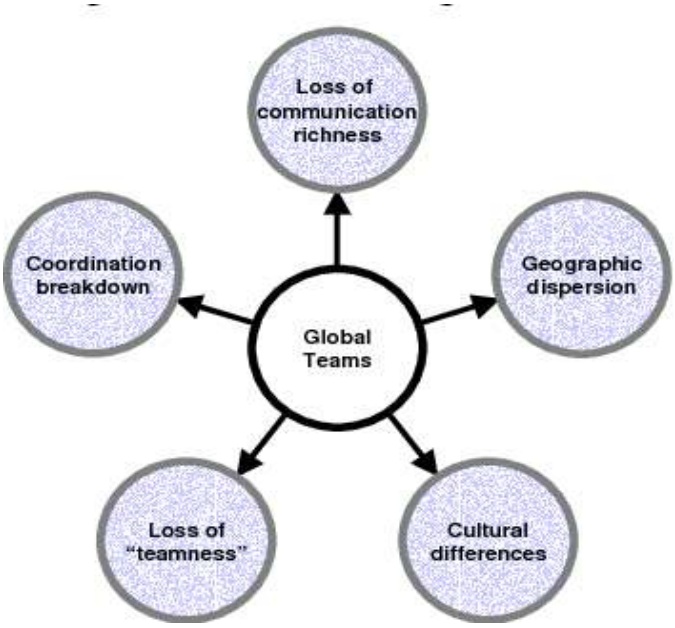


Fig. 1. Global Teams Issues.

Fig. 2. Agile Development.

This paper is structured as follows. Next, it offers a study of literature. Second, chaos and issues are discussed. Third, it presents and addresses the chaos and issues on communication in AGSD. Fourth, an overview of most discussed, less discussed chaos issues in the literature is presented. Then framework of GSD is elaborated. Determinately, it discusses future work and concludes the whole discussion in the last section

II. Literature Review

Agustin Yague and Juan Garbajosa conducted AGSD's Exploratory Communication Research. Instead, in this research work, observations were gathered from three points of view: contact between team members, communication about the status of the production process and the status of the progress of the product under development. The benefits of using media implementations have been explained by team members who assume that teams are co-located in honesty, such as perspicacious boards used by powerful video implements and the accumulation of media implements with centralized repository implements with process-generated data and product features that allow dispersed teams to share information. The results of exploratory research carried out on the effect of communication networks on AGSD are recorded in this paper. Due to distributed generated and traditional meetings suggested by flexible, often short in length and enhanced by face-to-face communication, interaction is critical in AGSD; thus, there are pellucid communication criteria for the amalgamation of the two [1]. The research analyzed the team's views of the facilities utilized for communication support. In this report, communication was addressed in three dimensions: communication between the team, the engineering process and the product under development [1].

Communication problems in the AGSD were identified by Nina Kamarina Kamaruddin, Noor Habibah Arshad and Azlinah Mohamed.

Fig. 3 described the agile methodology step by step. The goal is to resolve the issues associated with communication in Supple Ecumenical Software Development (AGSD). In order to assimilate knowledge from anterior works and from web reviews from practitioners worldwide, a literature survey was conducted. Utilizing a conceptual model, tabulated predicated on authors, and then addressed, the chaos issues are then illustrated. It is paramount to define the chaos issue in order to illustrate the authentic issues that subsisted in AGSD. It will somehow avail researchers and practitioners to identify the genuine quandaries that have arisen in AGSD, predicated on the established chaos quandaries cognate to communication. This paper introduces the chaos challenges encountered by members of the AGSD project team in communication based on existing literature as well as by global practitioners. All the problems that have really arisen in the AGSD setting are seen in the discussion. Our literature survey has established 13 communication-related chaos problems, with the key issues addressed in the literature as well as by the practitioners being various cultures and lack of regular face-to-face interaction [8]. In order to find more communication-related problems in AGSD, longitudinal studies will be carried out in future research.

Fig. 3. Agile Methodology.

Christof Ebert and Marco Kuhrmann in a manner to advance data and innovation move; it summed up points of view from the scholarly community and industry. It depended on an assessment of 10 years of examination and industry participation and experience announced at the IEEE International Conference on Software Engineering (ICGSE) arrangement. The aftereffects of their exploration show that GSE is a profoundly industry-attached field and, subsequently, an enormously goliath extent of ICGSE papers talk about the change of ideas and answers for programming to the ecumenical stage. Collaboration and teams, practices and organization, procurement and supply management, and performance factors were listed as the topics that engendered the most attention from researchers and practitioners. They also looked at emerging developments in GSE to promote more research and industrial cooperation beyond the study of past conferences. They summarized 10 years of ICGSE in their paper, and they searched for the issues explored in the past decade, cumulative information, and patterns. A discussion of the surviving state-of-the-art, focused on recently published studies and discussions by the different ICGSE conference committees, complemented their research. A analysis of the ICGSE papers showed that in the ICGSE conference series, both the conduct of high-quality research and the transition of subsisting software engineering concepts, procedures, practices and implements to the GSE context were addressed. This study has certain disadvantages that need to be discussed. In particular, the study at hand used well-known techniques to reuse validated relegation systems for secondary studies, and implemented systematic data collection and reporting processes. However, because the research focused exclusively on the ICGSE publication pool, they did not claim to show the full image, since they did not include additional publications, such as journal articles or conference papers published at other venues, in their report [6]. The source of their study in literature and only culled open discussions is another drawback. The notion of rigor, significance, and effect additionally affects this.

Yehia Ibrahim Alzoubi and Asif Qumer Gill adopted a SLR approach [4]. In the agile GSD background, and identified communication difficulties. Customized search and cull criteria for literature were first developed and then applied to relegate an accumulation of 449 initially documents. Lastly, for this review, 22 of 449 papers important to this research were chosen. These final 22 papers were analyzed and, in the sense of agile GSD, seven major categories of communication problems were identified. It is expected that the study results of their paper will assist researchers and practitioners to recognize agile GSD's communication challenges and develop methods, techniques and strategies to overcome these challenges. This study identified a range of problems to be tackled in order to build an efficient and effective agile framework GSD. The results of this research have been described in two steps. First, the accentuation of the research and the number of culled papers are registered. Secondly, it reports the information that was analyzed and interpreted from the culled studies to address the study concerns. This research helped us to change the current state of the art of versatile GSD communication issues. This research offers a knowledge base to agile practitioners and academics that have an interest in agile GSD. There are several disadvantages to this analysis, which are homogeneous to all other SLR studies. This paper is limited to the number of studies that have been checked from selected databases. There is a controversy about the use of an inhibited number of culled search databases and a finite number of search strings. This research accumulated papers from renowned databases, and we are completely confident that we have been provided with enough recent literature by the culled databases and search strings to review GSD's new agile communication challenges. Prejudice in the abolition of journals and inaccuracy in the extraction of information are the other paramount constraints of this SLR.

A. Research Questions

1) What are the chaos issues on Communication in AGSD?

2) What are the most discussed chaos issues on communication in AGSD in the literature?

3) What are the less discussed chaos issues on communication in AGSD in the literature?

III. Chaos and Issues

The study established a root concept (CATWOE) of the structure as visually perceived in order to explicate the "soft" quandary situation resulting from elements that perturb the stable environment of the engendered being studied [9]. Now, the Table I will show the root definition of the factors in the chaotic situation.

To semi-illuminate variables that lead to the creation of a chaotic situation in the creation of an involute system, Chaos theory is briefly demystified while contributors are engendered that lead to some instances of the chaotic situation. The principles of that system are the two key components of chaos theory: 1) Depend on an underlying order, and that 2) No matter how complex they may be. Very simple or small systems and operations can cause very complex behaviors or events. (The situation noted by Edward Lorenz, expressed as a delicate reliance on initial conditions, is this notion.)

TABLE I. Root Definition

C = Customer	Software development team, software owner, system user
A = Actors	Software development teams, project manager, system owner
T= Transformation	Chaos unwary unstable development environment to chaos ready a more stable development environment
W = developers who are aware of	Imminent change will be more prepared to navigate around a change than they who were caught unwary
O = Owner	System owner, government
E = Environment	Developers with the required know how, technology, method, and funds

IV. Chaos Issues on Communication in Agile Global Software Development

As verbally expressed in the precedent report, it can be considered a challenging task to handle challenges in the context of GSD by cumulating it with agile practices that further perplex things. Due to the distance involved, which somehow causes confusion in the creation process, these problems arise. This was accepted by the fact that the aforementioned uncertainty associated with AGSD problems in software development would have a negative impact on the software outcome. Communication is one of the main quandaries in Agile Global Software Development that has been highlighted in the literature. This difficulty can be considered a concern because it includes distance between locations and various time zones as well as different cultures. [10]. Hence, Fig. 4 will give a clear overview of chaos issues on communication in AGSD.

Nevertheless, communication is regarded as one of the essential elements of GSD, especially in a distributed agile environment where information sharing between team members is possible; understanding customer needs and development activities can be carried out efficiently and effectively [11].

We have listed thirteen issues related to the communication problem in AGSD from the literature. Fig. 6 depicts these problems. In highlighting the real issue that has really occurred in distributed agile programs, these problems are considered relevant.

A. Lack of Frequent Face-to-Face Contact

The biggest concern illustrated in the literature is the lack of face-to-face interaction. The explanation for this is that the distance from the venue and certain organizations allocated only a small portion of the foreign teams' travel budget. The development teams will try to meet during conferences, corporate training or workshops to connect face-to-face, plus personal meetings during personal meetings during holidays on rare occasions [18]. This problem resulted in the development team relying more on asynchronous communication as well as informal email communication as the right person is not accessible when required [19], [36].

B. Different Project Background

One of the quandaries illustrated in the literature is project history. Developers from various countries have various types of working culture, a comprehensive example linked to different project contexts, and this may lead to problematic quandaries when cross-border cooperation transpires [20]. Other than that communication turns out to be difficult for various interest groups to understand if agile approaches are new to the development teams involved, it takes time for development members to understand and information is important and should be conveyed to other developers as well as it takes time to change this culture because before that they used plan-driven development [18].

Fig. 4. Chaos Issues on Communication in AGSD.

C. Different Culture

One of the most important issues highlighted in literature as well as by practitioners is culture. Some of the cultural-related examples are that some of the members of the offshore team are hesitant to address negative or sensitive topics and only pass on positive data to the onshore team, cultural values and differences of ideology [18], [21], [22]. For example, the understanding of the culture and customs of the offshore teams related to festive seasons or holidays Cultural differences can affect team coordination and communication processes if not carefully handled [7], [8].

D. Different Language

Mundanely, distributed construction requires multiple locations or nations. When multiple team members from different countries and multiple members from different countries backgrounds and languages collaborate, it sometimes leads to great frustration. For example, offshore team members who are not native English verbalizers sometimes have arduousness interacting with native English verbalizers in English because of this; meeting often takes longer than mundane because it is arduous for them to communicate their conception [22], [23].

E. Low Quality of Telecommunication Bandwidth

Telecommunication bandwidth [18] another issue is one that needs to be answered. Often, via a communication medium, because of the context, tone and emotion were disoriented, an excess of time spent to describe things being addressed, and with poor quality of transmission hampering communication implements, communication networks can be slow and unreliable [8].

F. Misscommunication of Requirements

In the creation of software, specifications are considered to be the key component in the production of software that is functional and is continuously changing due to customer changes. If there is a lack of specific customer requirements data, it would have a bad effect on developers where developers have to come up with their own detailed requirements based on their past experience and try to understand what customers need at the same time [8].

G. Different Working Hours

Differences in time zones are one of the challenges that can be considered major and must be remembered. It improves the contact gap or overhead when there is distance in time zones. This resulted in the difficulties of arranging group meetings outside regular working hours at certain periods of the year due in particular during the winter due to the shorter time as well as the difficulty of holding long meeting instance of Sprint planning meeting [11].

H. Different Time Zones

One of the major challenges that need to be apperceived is the discrepancies in working hours. It raises the contact distance when there is a disparity in time zones, i.e. communication with the ecumenically dispersed team becomes arduous. In developed countries such as the USA, UK, etc. difficulties in consumer companies are typically expected, the time zone in distributed agile projects is normal [24].

I. Lack of Team Work

It is considered a team in distributed growth, even though it includes members of onshore and offshore teams. The software development team needs to collaborate and cooperate thoroughly in order to create a successful and quality product. Often difficulties arise when team members only connect with selected individuals in the team, such as only communicating with the Software Architect, contact is impacted because team members do not want to contribute to interacting with each other. The lack of structured contact can contribute to a decreased team spirit as well. There is also a problem in the growth of team spirit that is located, in two locations, or more particularly when communicating project priorities, goals and domain-specific as well as technical knowledge [8], [24].

J. Lack of Customer Involvement

Agile approaches are considered to rely more on people's communication than on engaging with clients in particular. The distance of the customers will usually be far away in the Agile distributed sense, resulting in the complexity of regular contact with them [24]. During the creation process, the customer did not offer complete commitment and often the partnership between the two parties is bad because they concentrate more on the process rather than individuals.

K. Unprepared Communication Tools

Contact is the predominant denotes of interaction For construction teams, both onshore and offshore, and with clients when it requires distance in space and time. It is important to have the right communication tools, but some organizations do not prepare teams with adequate and suitable communication tools, such as video conferencing or web-based conferencing facilities, especially when Scrum meetings are held [11]. It would make it hard to communicate efficiently without these facilities.

L. High Communication Cost

The least issues highlighted in the literature are the cost of communication impacting communication gaps [8], [24] where the cost of planning communication facilities is very high and companies need to prepare to spend money on it to provide an efficient means of communication between onshore, offshore and customers located in different locations.

M. Poor Communication Infrastucture

In order to encourage the distributed team to communicate with each other, communication infrastructure is considered essential and proper planning must be done. If this problem is not taken care of, there will be a lot of issues later on. For example, moving data to and sharing data with an offshore site typically reveals technical incompatibilities between sites [25]. The distribution of informal news or gossip during informal meetings, coffee breaks or after work meetings may somehow influence countries with poor infrastructure to prohibit rich conversations between team members.

V. An Overview of Most Discussed and Less Discussed Chaos Issues in the Literature

Table II demonstrates how much the literature discusses these problems.

With the aid of this table, we come to know that the problem is extreme in AGSD communication, i.e. developers and users also face them. We have come to realize that the communication difficulties that are addressed very few times in the literature are what. This gives researchers a new idea that this discussion or research is sufficient or we need to address these problems further because during AGSD they are often most trebly occurring problems. Fig. 5 highlights the mostly discussed communication chaos issues in AGSD.

TABLE II. Overview of Issues that have been Identified in Literature

ISSUES	LITERATURE
Lack of Frequent Face-To-Face Contact	[17],[18], [19], [25], [27], [28], [29],
Different Project Background	[8], [19],[26], [27]
Different Culture	[24], [25], [27], [28], [29],
Different Language	[8],[17], [24], [31]
Low quality of telecommunication background	[18], 28], [30]
Miscommunication of Requirements	[19], [32]
Different working hours	[8], [17], [27,[28], [29]
Different time zones	[11],[24], [30], [31], [34]
Lack of team work	[11], [25]
Lack of customer involvement	[29], [30], [33]
Unprepared communication tools	[11], [28], [30], [31]
High communication cost	[25], [28]
Poor communication infrastructure	[25], [28]

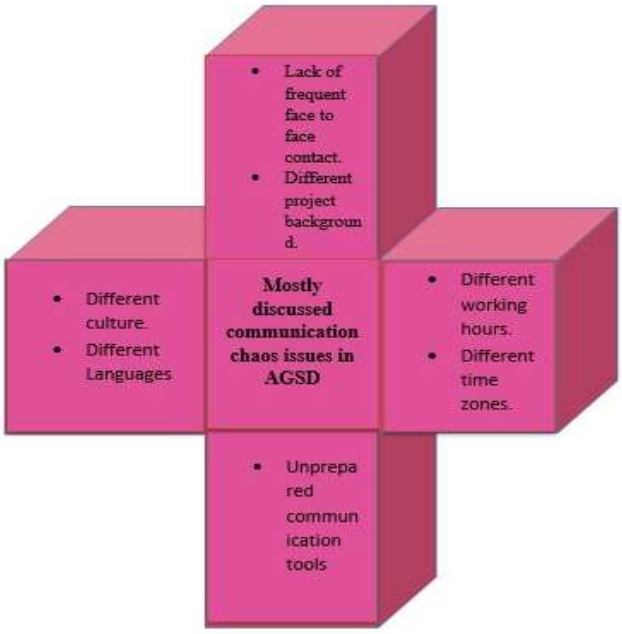


Fig. 5. Mostly Discussed Communication Chaos Issues in AGSD.

Based on the above statistics, there are two key problems that the researcher has regularly addressed and encountered by the practitioners. Different cultures and the lack of customary face-to-face interaction between scattered development teams are linked to the quandaries. Various operating hours or various time zones are also a problem often stated in the literature and by practitioners because of different geographical locations. This is accompanied by lack of confidence or ability to interact between team members, different histories of projects and different languages used between different countries. Fig. 6 highlights the less discussed chaos issues in AGSD.

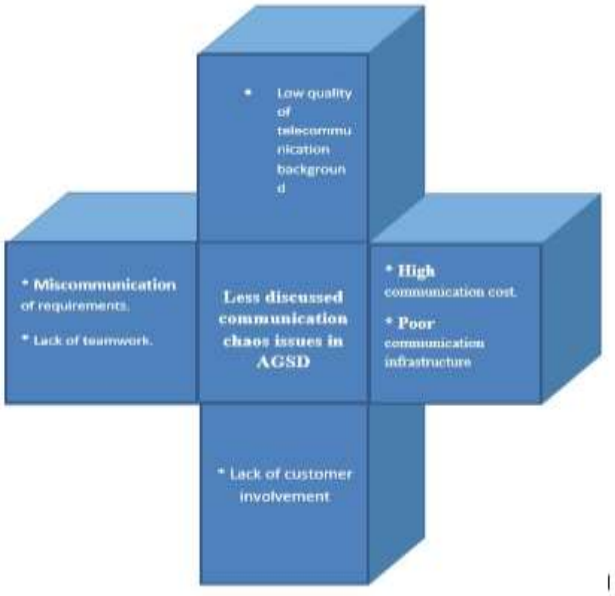


Fig. 6. Less Discussed Communication Chaos Issues in AGSD.

Lack of customer participation, efficiency of telecommunication capacity, communication costs, communication resources, lack of engagement or coordination by development teams, miscommunication of specifications and, last but not least, communication infrastructure or technological incompatibilities are the least listed problems that have been addressed.

VI. Framework of Global Software Developmet

Julia M. Kotlarsky and Jos van Hillegersberg [35] have researched and benchmarked (as a theoretical background) ecumenically distributed projects, and have developed an Ecumenical Software Development Project Structure that defines functional areas in which distributed teams cooperate during different project phases. During empirical data collection, they used these functional areas as a framework for group communication problems, teamwork and cooperation, and communication patterns and implements that we define. The framework describes and links different elements of projects in ecumenical software: product, methodology, project organization (i.e. individuals) and plans, and efficient planning, communication and control activities.

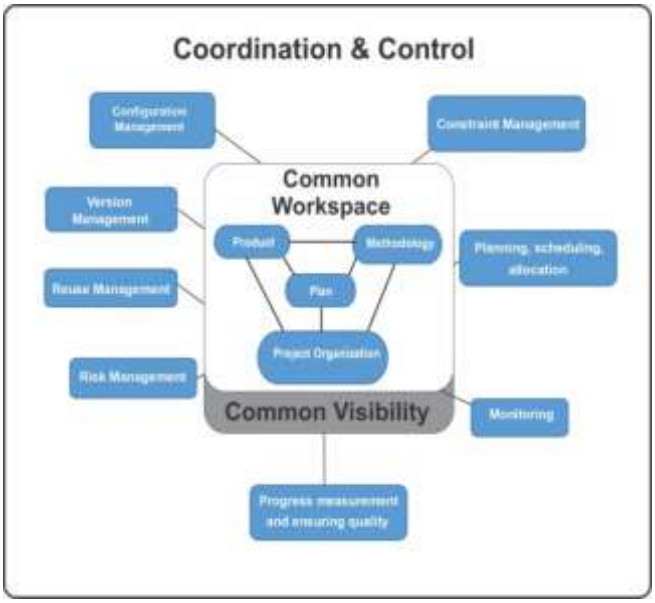


Fig. 7. Framework for Negotiation and Reaching of Consensus.

This framework shown in Fig. 7 will definitely help in the handling of chaotic situation in the agile GSD. So we can say that GSD teams should follow this framework in communication, coordination and control.

VII. Conclusion and Future Work

Over the last decade, global distribution for software production has become popular. There are a number of economic and technological trends that are expected to push distributed software development growth further. On the technical side, recent progress in information and communication systems makes it possible to work on software development projects in distributed modes through abstract distance perception.

In summary, this paper presents the chaos of communication based on the latest literature encountered by AGSD project team members. All the problems that have really arisen in the AGSD setting are seen in the discussion. Our literature survey found 13 communication-related chaos problems, with various cultures and lack of regular face-to-face interaction being the key issues addressed in the literature. The results of this review shows that there is need to be discuss these chaos issues more that are ignored by the researchers in the literature as these chaos issues are equally responsible in the formation of chaotic situation. At the end we also present a framework of coordination and control that will help to mitigate communication chaos issues in Agile GSD.

There is a desideratum in agile GSD to research communication issues and create or use implements, methods, and tactics to tackle them. Future studies would recommend the implementation in a distributed world of versatile approaches by designating timely implementing resources to alleviate these problems of chaos.

The results show that the key risk problem is communication in Agile Global Software development. The contributions of this paper may enable the scholars to propose and validate an enhanced approach to solve the problems of risk communication issues in the Agile GSD environment and assist practitioners in choosing the most effective and applicable method based on the Agile GSD project requirement. In the future, we will also concentrate on contributing to the field of understanding of software engineering management and software engineering models and methods.

REFERENCES

[1] Yagüe, Agustin, et al. "An exploratory study in communication in Agile Global Software Development." Computer Standards & Interfaces 48 (2016): 184-197.

[2] Jalali, Samireh, and Claes Wohlin. "Global software engineering and agile practices: a systematic review." Journal of software: Evolution and Process 24.6 (2012): 643-659.

[3] da Silva, Fabio QB, et al. "Challenges and solutions in distributed software development project management: A systematic literature review." 2010 5th IEEE International Conference on Global Software Engineering. IEEE, 2010.

[4] Alzoubi, Yehia Ibrahim, and Asif Qumer Gill. "Agile global software development communication challenges: A systematic review." Proceedings-Pacific Asia Conference on Information Systems, PACIS 2014. 2014.

[5] TaKeaways, Key. "The Forrester Wave: B2C Global Commerce Service Providers, Q1 2015." (2015).

[6] Ebert, Christof. Global software and IT: a guide to distributed development, projects, and outsourcing. John Wiley & Sons, 2011.

[7] Gill, Asif Qumer, Deborah Bunker, and Philip Seltsikas. "Evaluating a communication technology assessment tool (Ctat): a case of a cloud based communication tool." Proceedings-Pacific Asia Conference on Information Systems, PACIS 2012. 2012.

[8] Kamaruddin, Nina Kamarina, Noor Habibah Arshad, and Azlinah Mohamed. "Chaos issues on communication in agile global software development." 2012 IEEE Business, Engineering & Industrial Applications Colloquium (BEIAC). IEEE, 2012.

[9] Othman, Marini, Abdullah Mohd Zin, and Abdul Razak Hamdan. "Constructing a chaos proofing pre-development framework to manage chaos in a chaos-prone systems development environment." 2008 8th IEEE International Conference on Computer and Information Technology. IEEE, 2008.

[10] Jalali, Samireh, and Claes Wohlin. "Agile practices in global software engineering-A systematic map." 2010 5th IEEE International Conference on Global Software Engineering. IEEE, 2010.

[11] Dorairaj, Siva, James Noble, and Petra Malik. "Effective communication in distributed Agile software development teams." International Conference on Agile Software Development. Springer, Berlin, Heidelberg, 2011.

[12] Stray, Viktoria, and Nils Brede Moe. "Understanding coordination in global software engineering: A mixed-methods study on the use of meetings and Slack." Journal of Systems and Software 170 (2020): 110717.

[13] Khan, Arif Ali, and Muhammad Azeem Akbar. "Systematic literature review and empirical investigation of motivators for requirements change management process in global software development." Journal of Software: Evolution and Process 32.4 (2020): e2242.

[14] Barbosa, Hualter O., et al. "Developing a release management tool to support global software development: an experience report on Android platform." Proceedings of the 15th International Conference on Global Software Engineering. 2020.

[15] Camara, Rafael, et al. "Agile Global Software Development: A Systematic Literature Review." Proceedings of the 34th Brazilian Symposium on Software Engineering. 2020.

[16] Vallon, Raoul, et al. "Systematic literature review on agile practices in global software development." Information and Software Technology 96 (2018): 161-180.

[17] Podari, Zuriyaninatasa, et al. "Systematic Literature Review on Global Software Development Risks in Agile Methodology." 2020 8th International Conference on Information Technology and Multimedia (ICIMU). IEEE, 2020.

[18] Lee, Seiyoung, and Hwan-Seung Yong. "Distributed agile: project management in a global environment." Empirical Software Engineering 15.2 (2010): 204-217.

[19] Lehtonen, Ismo. "Communication challenges in agile global software development." University of Helsinki, Department of Computer Science, Faculty of Science (2009).

[20] Damian, Daniela, et al. "Awareness in the wild: Why communication breakdowns occur." International Conference on Global Software Engineering (ICGSE 2007). IEEE, 2007.

[21] ul Haq, Sami, et al. "Issues in global software development: A critical review." Journal of Software Engineering and Applications 4.10 (2011): 590.

[22] Marinho, Marcelo, Alexandre Luna, and Sarah Beecham. "Global software development: practices for cultural differences." International Conference on Product-Focused Software Process Improvement. Springer, Cham, 2018.

[23] Dorairaj, Siva, James Noble, and Petra Malik. "Understanding team dynamics in distributed Agile software development." International conference on agile software development. Springer, Berlin, Heidelberg, 2012.

[24] Kaur, Pawanpreet, and Sumit Sharma. "Agile software development in global software engineering." International Journal of Computer Applications 97.4 (2014).

[25] Kornstadt, Andreas, and Joachim Sauer. "Tackling offshore communication challenges with agile architecture-centric development." 2007 Working IEEE/IFIP Conference on Software Architecture (WICSA'07). IEEE, 2007.

[26] B. Lofland, iProject risk or issue?,h PM Technix. 12-2010.

[27] Bose, Indranil. "Lessons learned from distributed agile software projects: A case-based analysis." Communications of the Association for Information Systems 23.1 (2008): 34.

[28] Sauer, Joachim. "Agile practices in offshore outsourcing–an analysis of published experiences." Proceedings of the 29th information systems research seminar in Scandinavia, IRIS. Vol. 29. 2006.

[29] Dullemond, Kevin, Ben van Gameren, and Rini van Solingen. "How technological support can enable advantages of agile software development in a GSE setting." 2009 Fourth IEEE International Conference on Global Software Engineering. IEEE, 2009.

[30] Hossain, Emam, Muhammad Ali Babar, and Hye-young Paik. "Using scrum in global software development: a systematic literature review." 2009 Fourth IEEE International Conference on Global Software Engineering. Ieee, 2009.

[31] I would like to know what are the issues/challenges/problems that exist in Agile Global Software Development. Care to share your experiences?,h Agile Alliance. Dec-2010.

[32] Korkala, Mikko, Minna Pikkarainen, and Kieran Conboy. "Distributed agile development: A case study of customer communication challenges." International Conference on Agile Processes and Extreme Programming in Software Engineering. Springer, Berlin, Heidelberg, 2009.

[33] Korkala, Mikko, and Pekka Abrahamsson. "Communication in distributed agile development: A case study." 33rd EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO 2007). IEEE, 2007.

[34] Layman, Lucas, et al. "Essential communication practices for Extreme Programming in a global software development team." Information and software technology 48.9 (2006): 781-794.

[35] Kotlarsky, Julia M., Kuldeep Kumar, and Jv Hillegersberg. "Coordination and collaboration for globally distributed teams: the case of componentbased/object-oriented software development." Proceedings of International Workshop on Global Software Development (ICSE 2002). 2002.

[36] Shameem, Mohammad, et al. "Taxonomical classification of barriers for scaling agile methods in global software development environment using fuzzy analytic hierarchy process." Applied Soft Computing 90 (2020): 106122.

[37] Sinha, Richa, Mohammad Shameem, and Chiranjeev Kumar. "SWOT: strength, weaknesses, opportunities, and threats for scaling agile methods in global software development." Proceedings of the 13th innovations in software engineering conference on formerly known as India software engineering conference. 2020.

Urban Addressing Practices and Geocoding Algorithm Validity in Developing Countries

Case of Casablanca City - Morocco

Mohamed El Imame MALAAININE[1]
EHTP
Casablanca-Morocco

Hatim LECHGAR[2]
FSAC-Hassan II University
Casablanca-Morocco

***Abstract*—Addressing systems have a key role in understanding and managing economic connections and social conditions, especially in urban territories. Developing countries need to learn from previous experiences and adapt solutions and techniques to their local contexts. A review of the world bank's experience in addressing cities in Africa during the 1990s provides valuable lessons. It provides an understanding of the operational issues and the key success factors of such operations. It also helps to understand the conceptual components of these systems and the efforts required to build them in the field before the creation of their IT infrastructure. An addressing experience from a private sector initiative in Casablanca-Morocco is also reviewed, where efforts concern the creation of a comprehensive database of addresses. The methods used to collect the data in the field are presented as well as the conceptual model for its integration. The validity of geocoding techniques, which represent the core computing tools of addressing systems, is discussed. In the Moroccan context, the official addressing rules follow Western models and standards, used by default in geocoding algorithms. The study of data collected in Casablanca, processed with GIS tools and algorithms, shows that the percentage of cases not respecting these rules is far from negligible. The analysis was particularly interested in the two main criteria of address numbers: "parity" and "respect of intervals", analyzed by street segment. Compliance with these conditions was only observed at about 53%. It is then concluded that a geocoding system based on a linear model is not sufficiently validated in the Moroccan context.**

***Keywords*—Addressing system; geocoding; Geographic Information System (GIS)**

I. Introduction

Addresses are necessary data for citizens, administrations and companies. Through an address, a citizen can have access to several civil rights and public services; Administrations can efficiently manage their territories and companies can manage and optimize supply chains. It was once believed that about 80% of information, especially those used by local authorities, have a geographical component, related in a way or another to address locators [1], [2]. In the times of IoT (Internet of Things), it's hardly possible today to find data without spatial coordinates. While the latest geocoding literature deals with the latest techniques in the matter, such as machine learning [3] and deep learning particularly [4], the classical issues related to historical address structure and standardization remain relevant [5,6]. The general literature deals with geographic related applications in different countries such as in Australia [7], Brazil [8], China [9], Croatia [10], Cuba [11], Germany [12], India [13], Morocco [14], Quebec [15], South Africa [16], Turkey [17], etc. The applications based on address locators are more than ever evolving, and the need for reliable and accurate address systems has never been more. Unfortunately, while such systems have already reached the stage of maturity in developed parts of the world [18], [19], it remains a real issue in developing countries. However, it is there where it's the most needed, for basic applications, already discussed in research works in other contexts, such as health studies [20]-[22], politics [23], criminality [24], traffic accidents [25], emergency dispatching [26], etc.

In developing countries, the issue of addressing systems presents a big challenge, including norms on the field, availability and quality of the reference data and reliability of geocoding techniques. On the field, addresses numbers and streets names should be assigned according to logical and consistent methods. The quality of geocoding, which consists of transforming a given number of descriptive into a geographic position [27], will then depends on the quality of both the reference database of addresses and the used methods.

II. Review of Urban Addressing in Africa and Morocco

A. The World Bank Experience in Africa

The addressing process is a critical issue for the city. It is a challenge to be taken up by several stakeholders, including town planners who plan the base of future addresses, local authorities who assign formalized addresses, install and maintain signs for street names and squares, utilities who use addresses when providing services or billing, postal operators who deliver mail to an address, as well as residents who maintain the numbering plates of their buildings and can correct errors in their addresses.

This complex operation includes the formalization of the rules of reference for the addressing process. It consists of creating and updating standardized addresses in the city. The two important operations carried out on the field are: the naming of the streets and the numbering of the buildings.

TABLE I. THE TECHNICAL FEATURES OF ADDRESSING PRACTICES IN SOME AFRICAN COUNTRIES

	Burkina-Faso	**Cameroun**	**Guinea**	**Niger**
Division	sectors	zones	municipalities	neighborhood groups
Street codification	sector & order number	zone & order number	neighborhood initials & order number	neighborhood initials & order number
Numbering doorways	metric and alternating	metric and alternating	metric and alternating	metric and alternating
street signs supply	international bidding process	local firm (bidding process)	local firm (bidding process)	international bidding process
Street sign installation	small local companies	municipal employees & NGO survey takers	addressing unit and municipal technical departments	local company under supervision of addressing unit
Surveys and number assignment	addressing unit		addressing unit	addressing unit
Survey data	address occupant's name, plot use category, and cadastral references	address, occupant's name, plot use category, cadastral references, type of activity	address, occupant's name, plot use category, water and electricity meter numbers.	address, occupant's name, plot use category, the cadastral reference, whether or not electricity and water are available, and information on streets
Address directory	specialized software program	addressing Software developed locally	special software program	address management software program

The World Bank carried out several addressing experiments during the 1990s, in different African cities [28]. Table I presents an overview of the technical features of addressing practices in Burkina-Faso (Ouagadougou and Bobo-Dioulasso), Cameroun (Yaoundé and Douala), Guinea (Conakry) and Niger (Niamey). The world bank's financially and technically supported addressing projects extended to several other African cities in Mali, Mauritania, Mozambique, Senegal, Benin, Rwanda, Djibouti, Togo, and Côte d'Ivoire (Ivory Coast). Fig. 1 shows the concerned countries.

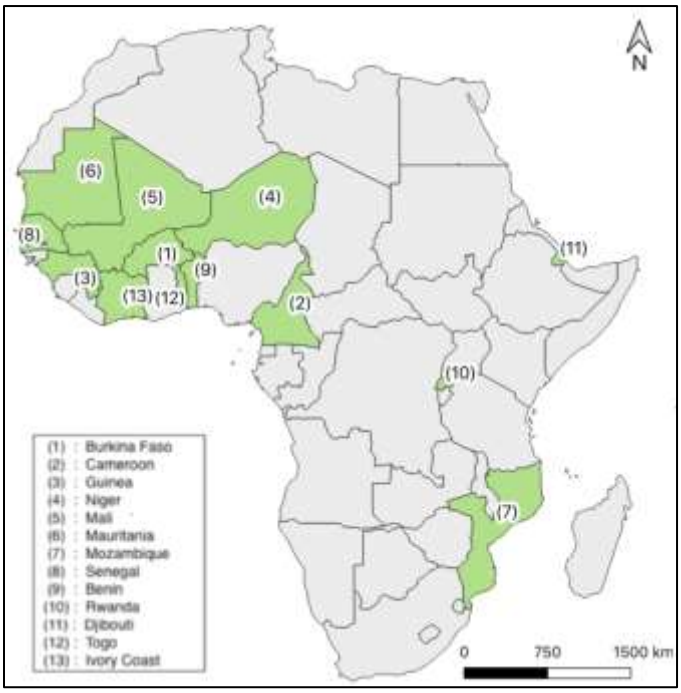


Fig. 1. Addressing Projects Countries, Supported by the World Bank during the 1990s.

The World Bank's recommendations for addressing operations are based on the observation that it is almost impossible to name all streets that addressing operations are - first of all - a municipal action and that addresses are to be defined in relation to the streets and not in relation to the blocks.

The key success factors for successful addressing operations that were concluded from these experiences are: organization and motivation of the addressing unit; the involvement of the municipalities, decision-makers and technical services (which must have the necessary means and skills); financial efficiency during the project (while having good control); the simplicity of the database and the software developed to facilitate transition after the project phase (in particular to ensure that addresses are updated); controlling the scope of the project (concentration of efforts on the pure and simple objective of addressing); good coordination with stakeholders, in particular utilities and the post offices.

The main indicators that were used to assess the outcome of these projects are: the budget of the operation, the number of street signs installed, the number of buildings "addressed" (percentage of households concerned) and cost per capita and per addressed door. In the long term, the growth rates of local services such as tax collection and postal services should confirm the success of these operations.

B. Private Sector Initiative in Casablanca-Morocco

The first known addressing project, aiming to create a comprehensive database of addresses, inventorying all address locators of a major Moroccan city was initiated in Casablanca city in the late 2000s. It is the private company, insuring the delegated management of water and electricity utilities in the Grand Casablanca that was behind this initiative. In the absence of providers of such important data, critical for its operations, the company had to collect more than 400000 address locators. Fig. 2 shows the projects area.

Fig. 2. Projects Area of the Private Sector Initiative in Casablanca-Morocco.

After planning and reference data acquisition, such as streets and necessary base map data, 200 operators were sent to the field, equipped with 3500 printed plans in total. Each map concerns a specific sector and contains the reference data necessary to recognize address locators to collect: streets, plots, remarkable places, neighborhoods and sector limits [29].

The targeted area is divided into sectors, well known and mastered by the field operators. Each sector is printed in suitable format map, with the sector reference included. Once the field survey is performed, with both positions and descriptive information required for the matching with the company's operational database, the maps are handed to the back-office for filling the project database. A sub set of these maps is then used for quality control. All maps are in the end scanned and archived.

C. Discussion of Addressing Field Operations Practices

In order to create an addressing database (as in the case of the private initiative in Casablanca), or to prepare an inventory to improve addressing in the city (as in the case of initiatives assisted by the World Bank in Africa), complex field surveys operations are required. Two missions can be distinguished, to be carried out successively, since the first one makes it possible to better prepare the second one by providing its necessary reference data. First, mission I, the survey of streets and their signs, then mission II, the survey of the numbering of buildings.

Before starting the field operations, data model conception of information to be collected is necessary. Table II shows the key data for the two missions.

Once identified, these data should be integrated into a larger addressing data model, such as the one presented in Fig. 3 using the UML formalism.

1) Mission I: Streets survey: Given that the area to be covered is often very large, (for instance, more than 1220 km2 in the case of the Casablanca project), the field surveys must be organized by geographic elements that can be mastered and easily navigable, in order to allow fluidity of operations on the field. All data that can be acquired before the field mission, such as the streets plots, must be integrated beforehand, so that the strict minimum data should be gathered from the field.

The choice of street segments as survey basic elements has the following advantages: allow the control of the total coverage of the study area; optimize the circuits of passage in the field; allow the allocation of different sections, belonging to the same street, to several operators (which corresponds to the logic of the administrative division of cities); gather information that may change from one street segment to another, such as width; prepare for the city signs plan which is designed by segment of street.

2) Mission II: Numbering survey: For the same reason of optimization, the use of streets' segments as basic elements for organizing this second mission remains relevant. It would also be possible to combine this logic with an administrative or business division of the project area, in order to define circuits that are easily recognizable by the field operators.

Another reason to consider the streets segments for this second mission too would be to anticipate the preparation of basic data for the development of a geocoder, the quality of which improves while using street segments instead of streets.

TABLE II. KEY ADDRESSING DATA

	Entity	Key attributes
Mission I	Streets	Geometry (linear), name, type, width, length
	Streets signs	Geometry (Point), type (plate or panel), text, condition
Mission II	Address locations	Geometry (Point), Number in the road, Name of the road, Name of building), Type (Building, villa, house…), Use (Residential, commercial, industrial, mixed), Activities

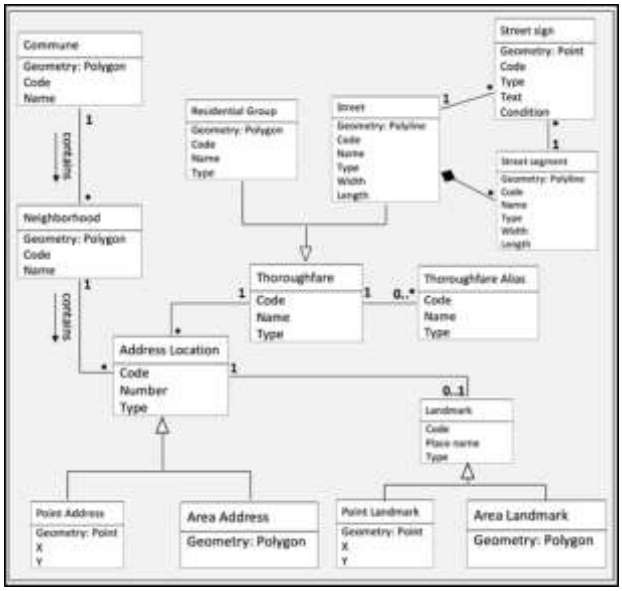


Fig. 3. UML Addressing Data Model.

III. Review of Geocoding Techniques and Conditions

A. Geocoding Techniques

The principle of geocoding methods is to compare a list of descriptive address elements with a well-structured reference database. This operation is done in three steps that represent the general geocoding algorithm, which are well documented in the literature [30]-[32], shown in the Fig. 4.

The Geocoding process manipulates "Input data" in order to get "Output data" through a "Matching algorithms" using a structured "Reference database"; those are the four parameters of geocoding and here after their dynamics. In Input, data to geocode is introduced, in form of a list of descriptions such as a postal address details. In Output, a georeferenced data is returned, with the geometry that is supported by the processing algorithm. It is often a two-dimensional point, but it can also be another form of complex data such as 3D objects [33]. It is the matching algorithm that decides of the corresponding result, based on the input data, the reference data and matching rules. This is why research works focuses especially on matching algorithms.

When all the address points' locations are available, in the case of a comprehensive reference database, the matching operation becomes quite evident. A simple search request is enough to find the exact coordinates to return. On the other hand, in the case of linear alternating numbering model, the returned position is rather calculated. It is an interpolated position based on the elements available in the linear referencing database, notably in the streets and street segments, such as intervals and parity of numbers on each side of the streets segment [34], [35].

B. Geocoding Algorithm Preconditions

The geocoding algorithms, in the case of linear alternating numbering model widely prevailing, begins with determining the street segment which corresponds to the descriptive list of the searched location. This first step uses the address numbers interval in the attributes of the street segments. Once the segment is determined, the address side (right or left) is concluded from the parity types (odd or even) of the numbers on either side of the segment, and the parity of the number in the searched address location. The exact position on the corresponding side is then calculated. Other parameters such as distance and angle from the segment and offsets from its ends can fine-tune the accuracy of the estimated position.

Other data can also improve this accuracy, such as information on the number of buildings on each side and their geographical distribution [36],[37].

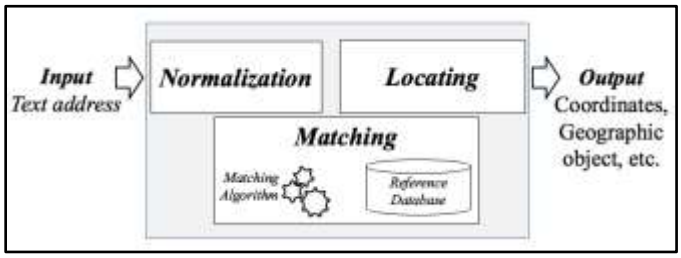


Fig. 4. Geocoding General Process.

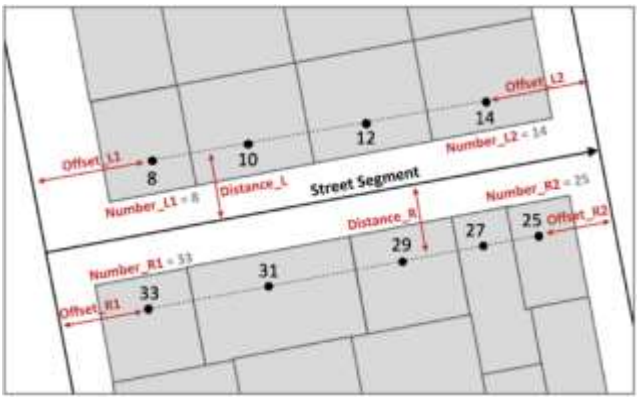


Fig. 5. Geocoding Parameters Overiew.

Fig. 5 shows an example of Geocoding parameters in the case of linear alternating numbering, where Number_L1 and Number_L2 are first and last numbers on the left side; Number_R1 and Number_R2 are first and last numbers on the right side; Offset_L1, Offset_L2, Offset_R1 and Offset_R2 are the offsets from the left and right ends; Distance_L and Distance_R are the distances from the segment.

For the geocoding algorithm to be effective, two main prerequisites must be satisfied by addresses in the field: consistent number intervals per street segment and consistent parity on each side. The uniformity of the distribution of buildings on each side of the street segment improves the accuracy of the calculated position.

IV. Methodology

In this section, the evaluation of the main prerequisites of geocoding is studied, notably address numbers parity and intervals, in the city of Casablanca.

A. Study Area and used Data

The main data used are the streets shapefile and address locators collected in the field as part of the private initiative project in Casablanca, presented in Section II. This database of 63,833 address locators of the communes of: Anfa, Maarif, Mers Sultan, Sidi Belyout and El Fida, represents approximately 15% of the total number of address locators in Grand Casablanca (Fig. 6).

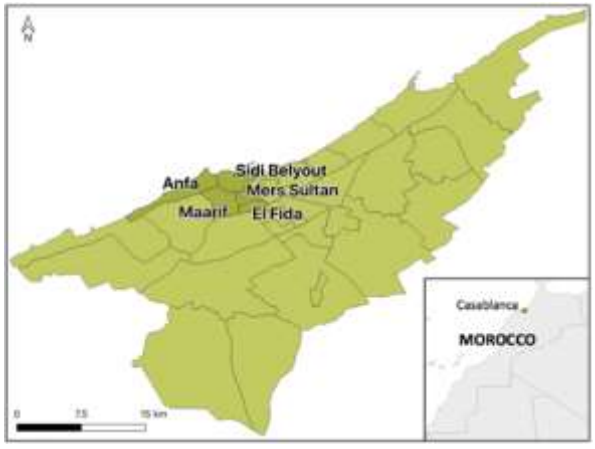


Fig. 6. Case Study Area.

From the streets shapefile, using a GIS software, street segments are generated. Then, for each address point, the relative position (left or right) with respect to its street segment is calculated. Finally, for each street segment, the geocoding parameters are calculated.

B. Geocoding Preconditions Analysis Method

To assess the validity of the main prerequisites of geocoding presented previously, two main questions need to be answered: For what percentage, the numbers of address locators are consistent with the parity condition of their side of the street segment? And what percentage of the address locators' numbers fall within the range of the numbers' interval on their side of the street segment?

For address locators belonging to a given street segment S, let:

- PS (OL): The percentage of address locators with an Odd number that are on the Left side of the street segment.
- PS (OR): The percentage of address locators with an Odd number that are on the Right side of the street segment.
- PS (EL): The percentage of address locators with an Even number that are on the Left side of the street segment.
- PS (ER): The percentage of address locators with an Even number that are on the Right side of the street segment.

For an odd number and an even number belonging to a street segment S, the probability that the even number is on the left side and the odd number is on the right is:

$$PS\ (EL \text{ and } OR) = P(EL) \times P(OR) \quad (1)$$

Similarly:

$$PS\ (ER \text{ and } OL) = P(EL) \times P(OR) \quad (2)$$

Since the the street segments that meet the parity condition are those that have all odd numbers on one side and all even numbers on the other side, these streets are those verifying:

$$P\ (EL \text{ and } OR) = 1 \text{ or } P\ (ER \text{ and } OL) = 1 \quad (3)$$

Among the address locators belonging to such a street segment, those which meet the number range condition are those whose numbers are exclusively within the range of this street segment (taking into account all segments belonging to the address's street).

V. Results Discussion and Conclusion

Table III shows that 87% of street segments verify (3), then fulfil the first condition: the consistency of the parity of the address numbers. The percentage of address locators belonging to these streets segments is 83% as shown in Table IV. Among these points, 61% have a number that belongs to one and only one range of segment numbers, for this also meet the second condition: the consistency of the rages of numbers (Table V).

TABLE III. Segments Parity Consistency

Case	Number of segments	Percentage
Even numbers on the right & Odd numbers on the left	3,472	47%
Even numbers on the left & Odd numbers on the right	2,885	39%
Even numbers on one side & odd numbers on the other	6,357	**87%**
All street segments	7,325	100%

TABLE IV. Address Locators Parity Consistency

Case	Number of points	Percentage
Even numbers on the right & Odd numbers on the left	27,062	45%
Even numbers on the left & Odd numbers on the right	22,269	37%
Even numbers on one side & odd numbers on the other	49,331	**83%**
All Points (having a street as toponym)	59,741	100%

TABLE V. Address Locators Intervals Consistency

Case	Number of points	Percentage
In one interval	29,892	**61%**
In many intervals	19,439	39%
All points (parity consistent)	49,331	100%

If we consider all the address locators with a street toponym in the study area, the percentage of compliance with the two rules is around 53%. This percentage drops to only 47%, if we consider all address locators, regardless of the types of toponyms, since 6% of address locators are not linked to a street toponym.

These results indicate that the conditions necessary for the use of standard geocoding, based on linear alternating numbering model according to Western standards, are not met in the Moroccan case. Thus, the establishment of a national addressing system based on geocoding could not ensure the quality necessary for the applications which depend on it.

This problem can be explained by the lack of application of addressing standards in the field: temporary addresses allocated to housing development projects that become permanent, streets that are not assigned official names for long periods of time (years in some cases), addressing services which lack resources and coordination in cities experiencing a rapid expansion which further complicates the situation.

That said, the bulk of the problem is first organizational before it is technical. So, in order to be able to set up a reliable reference addressing system, urgent solutions must be proposed and others must be established over time.

Thus, in the short term, addressing campaigns like those of the World Bank in Africa must be carried out to overcome the lack of address references in the field and to promote the addressing of cities. Such campaigns will also make it possible to have up-to-date and more general data on the addressing situation throughout the country, the latest data available only

concerning the city of Casablanca and already dating back almost ten years. Addressing information systems must be built around comprehensive databases and not on interpolation algorithms as long as the addresses in the field does not comply with the standards.

Over time, the addressing standards themselves as well as the address attribution procedures must be improved. They must be integrated into the urbanization and management processes of the city since early stages in order to avoid temporary, non-compliant solutions that last. This surely cannot be done without a good governance, led by specialized organizations which collaborate with businesses and research institutions.

REFERENCES

[1] Davis, C. A., & Fonseca, F. T. (1993, July). Address base creation using raster/vector integra-tion. In Papers from the Annual Conference-Urban and Regional Information Systems Association (PP. 39-39). Urisa Urban and Regional Information Systems.

[2] Eichelberger, P. (1993). The importance of addresses-the locus of gis. In Papers from the Annual Conference-Urban and Regional Information Systems Association (PP. 212-212). Urisa Urban and Regional Information Systems.

[3] Lee, K., Claridades, A. R. C., & Lee, J. (2020). Improving a Street-Based Geocoding Algorithm Using Machine Learning Techniques. Applied Sciences, 10(16), 5628.

[4] Yin, Z., Ma, A., & Goldberg, D. W. (2019). A deep learning approach for rooftop ge-ocoding. Transactions in GIS, 23(3), 495-514.

[5] Kirielle, N., Christen, P., & Ranbaduge, T. (2019, December). Outlier Detection Based Accurate Geocoding of Historical Addresses. In Australasian Conference on Data Mining (pp. 41-53). Springer, Singapore.

[6] Matci, D. K., & Avdan, U. (2018). Address standardization using the natural language process for improving geocoding results. Computers, Environment and Urban Systems, 70, 1-8.

[7] Christen, P., Churches, T., & Willmore, A. (2004, December). A probabilistic geocoding system based on a national address file. In Proceedings of the 3rd Australasian Data Mining Confer-ence, Cairns.

[8] Davis Jr, C. A., & de Alencar, R. O. (2011). Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city. Transactions in GIS, 15(6), 851-868.

[9] Tian, Q., Ren, F., Hu, T., Liu, J., Li, R., & Du, Q. (2016). Using an optimized Chinese address matching method to develop a geocoding service: a case study of Shenzhen, China. ISPRS In-ternational Journal of Geo-Information, 5(5), 65.

[10] Cetl, V., Kliment, T., & Jogun, T. (2018). A comparison of address geocoding techniques–case study of the city of Zagreb, Croatia. Survey Review, 50(359), 97-106.

[11] de Armas García, C. J., & Cruz Gutiérrez, A. A. (2013). Deployment of a National Geocoding Service: Cuban Experience. Journal of the Urban & Regional Information Systems Associa-tion, 25(1).

[12] Ahlers, D., & Boll, S. (2009). On the accuracy of online geocoders. Geoinformatik 2009.

[13] Chatterjee, A., Anjaria, J., Roy, S., Ganguli, A., & Seal, K. (2016, October). SAGEL: smart ad-dress geocoding engine for supply-chain logistics. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 1-10).

[14] Malaainine, M. E. I., Rhinane, H., Baidder, L., & Lechgar, H. (2013). Omt-g modeling and cloud implementation of a reference database of addressing in Morocco.

[15] Burns, S., Miranda-Moreno, L., Stipancic, J., Saunier, N., & Ismail, K. (2014). Accessible and practical geocoding method for traffic collision record mapping: Quebec, Canada, case study. Transportation research record, 2460(1), 39-46.

[16] Coetzee, S., & Cooper, A. K. (2007). What is an address in South Africa?. South African Journal of Science, 103(11-12), 449-458.

[17] Yildirim, V., Yomralioglu, T., Nisanci, R., & Inan, H. (2014, June). Turkish street addressing system and geocoding challenges. In Proceedings of the Institution of Civil Engi-neers-Municipal Engineer (Vol. 167, No. 2, pp. 99-107). Thomas Telford Ltd.

[18] Martin, D. (1999). Spatial representation: the social scientist's perspective. Geographical in-formation systems, 1, 71-80.

[19] Boscoe, F. P., Ward, M. H., & Reynolds, P. (2004). Current practices in spatial analysis of can-cer data: data characteristics and data sources for geographic studies of cancer. International journal of health geographics, 3(1), 28.

[20] Cromley, E. K. (2019). Using GIS to address epidemiologic research questions. Current Epi-demiology Reports, 6(2), 162-173.

[21] Boulos, M. N. K. (2004). Towards evidence-based, GIS-driven national spatial health infor-mation infrastructure and surveillance services in the United Kingdom. International Journal of Health Geographics, 3(1), 1.

[22] Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., & Zimmer-man, D. L. (2006). Geocoding in cancer research: a review. American journal of preventive medicine, 30(2), S16-S24.

[23] Haspel, M., & Knotts, H. G. (2005). Location, location, location: Precinct placement and the costs of voting. The Journal of Politics, 67(2), 560-573.

[24] Ratcliffe, J. H. (2001). On the accuracy of TIGER-type geocoded address data in relation to ca-dastral and census areal units. International Journal of Geographical Information Sci-ence, 15(5), 473-485.

[25] Qin, X., Parker, S., Liu, Y., Graettinger, A. J., & Forde, S. (2013). Intelligent geocoding system to locate traffic crashes. Accident Analysis & Prevention, 50, 1034-1041.

[26] Zimmerman, D. L., Fang, X., Mazumdar, S., & Rushton, G. (2007). Modeling the probability distribution of positional errors incurred by residential address geocoding. International Journal of Health Geographics, 6(1), 1.

[27] Hutchinson, M., & Veenendall, B. (2005, August). Towards using intelligence to move from geocoding to geolocating. In Proceedings of the 7th Annual URISA GIS in Addressing Con-ference.

[28] Farvacque-Vitkovic, C., Godin, L., Leroux, H., Chavez, R., & Verdet, F. (2005). Street address-ing and the management of cities. The World Bank.

[29] Malaainine, M., Slaoui, L., Rhinane, H., & Baidder, L. (2010, May). Méthode de Saisie en Masse des Points d'Adresses Géolocalisés de la Grande Ville de Casablanca. In First International Congress on GIS and Land management (SIGGT 2010), Casablanca (p. 101).

[30] Goldberg, D. W. (2009). A Geocoding Best Practices Guide, The North American Association of Central Cancer Registries. http://www. naaccr. org/filesystem/pdf/Geocoding_Best_Practices. pdf.

[31] Karimi, H. A., Sharker, M. H., & Roongpiboonsopit, D. (2011). Geocoding recommender: an algorithm to recommend optimal online geocoding services for applications. Transactions in GIS, 15(6), 869-886.

[32] Hutchinson, M. J. (2010). Developing an agent-based framework for intelligent ge-ocoding (Doctoral dissertation, Curtin University).

[33] Lee, J. (2004, October). 3D GIS for geo-coding human activity in micro-scale urban environ-ments. In International Conference on Geographic Information Science (pp. 162-178). Spring-er, Berlin, Heidelberg.

[34] Davis Jr, C. A., Fonseca, F. T., & Borges, K. A. (2003, October). A Flexible Addressing System for Approximate Geocoding. In GeoInfo.

[35] Cayo, M. R., & Talbot, T. O. (2003). Positional error in automated geocoding of residential ad-dresses. International journal of health geographics, 2(1), 10.

[36] Bakshi, R., Knoblock, C. A., & Thakkar, S. (2004, November). Exploiting online sources to ac-curately geocode addresses. In Proceedings of the 12th annual ACM international workshop on Geographic information systems (pp. 194-203).

[37] Wu, J., Funk, T. H., Lurmann, F. W., & Winer, A. M. (2005). Improving spatial accuracy of roadway networks and geocoded addresses. Transactions in GIS, 9(4), 585-601.

A Self Supervised Defending Mechanism Against Adversarial Iris Attacks based on Wavelet Transform

Meenakshi K[1]
Department of Computer Science and Engineering
School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

G. Maragatham[2]
Department of Information Technology
School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

***Abstract*—In biometric applications, deep neural networks have presented significant improvements. However, when presenting carefully designed input training data known as adversarial examples, their output is severely reduced. These types of attacks are termed as adversarial attacks, and any biometric security system is greatly affected by these attacks. In the proposed work, an effective defensive mechanism has been developed against adversarial attacks which are introduced in iris images. The proposed defensive mechanism is following the concept of wavelet domain processing and it investigates the mid and high frequency components of wavelet domain components. Based on this, the model reproduces the various denoised copies of input iris images. The proposed strategies are intended to denoise each sub-band of the wavelet domain and assess the sub-bands most likely to be affected by the adversary using the reconstruction error measured for each sub-band. We test the effectiveness of the proposed adversarial protection mechanism against various attack methods and analyzed the results with other state of the art defense approaches.**

***Keywords*—Iris classification; deep neural networks; adversarial attack; defense method; wavelet processing; biometrics**

I. Introduction

In various classification applications such as biometric spam filtering, autonomous vehicle system, and speech recognition, etc. machine learning and deep learning-based classifiers have now achieved outstanding performance [1, 2]. Besides that, the classifiers are more prone to adversarial attacks that make the classification models behave more confidently in the wrong direction, i.e. the model misclassified the sample. Adversarial examples have been created by these attacks that are classified as data samples built to manipulate the Classifier model [3] The adversary has thus used these adversarial examples and these compromised examples to target the security system to give access to unauthorised users in order to modify the identity of the actual subject. So the adversarial examples are considered as security risks which are structured to affect the performance of the ML based classifier [4]. The adversarial attacks are classified as two types: 1. Black Box and 2. White box attacks. The attacker has a detailed understanding of the layout of the classification model in the case of a white box attack, such as parameters and algorithms used, etc. [5], whereas the adversary has no knowledge about the classification model in the black box method [6]. The techniques for coping with adversarial threats are called defensive methods. The defence mechanisms focus on making the classification model safer and more stable, and few methods seek to recognise the adversarial data, i.e. manipulated image [7]. To identify the person uniquely, various biometric characteristics such as fingerprint, face, iris, signature, voice, retina etc. are used. In [8], the important features of iris are captured which makes it more significant and secure biometric trait for the unique identification of an individual with a high degree of confidence. But the attacker introduces adversarial examples (manipulated iris images) to fool the recognition system and it is a big challenge to the security system. Protecting the iris recognition system from these types of attacks is important and it is a significant research direction to define the necessary countermeasures used to effectively detect adversarial attacks.

The Wavelet Decomposition technique is used in the proposed paper to classify the manipulated adversarial data. Kim et al. [9] have already shown that wavelet components of iris image with low and low-mid frequencies have high data to detect the subject, and these components are reliable and difficult to inject noise. To build the manipulated samples, the adversaries add the high frequency sections to the iris images. On the basis of this fact, we have proposed a defensive mechanism that effectively recognises adversarial attacks.

The following contributions are presented in this paper: a) An efficient defensive mechanism has been implemented which is applied before the iris recognition process. b) The proposed work analyzes the wavelet components, to identify the adversarial data and it is accurate and stable against adversarial attacks. C) The proposed methodology is compared with other state of the art defensive mechanisms in terms of accuracy.

The paper is organized in the following way. Related works are presented in Section II. Section III explains in depth the proposed approach. The experimental results are listed in Section IV. The conclusion and possible future developments are drawn in Section V.

II. Related Works

A. Adversarial Attacks

Recently, deep learning models have performed tremendously in a large range of applications like biometrics [10, 11], security [12, 13], autonomous vehicle control systems and Spam Filtering. However these models are more susceptible to manipulated input data which is called adversarial examples. Synthetic information is described as the small disturbances are added to the input image, often referred

to as poisoning data. It has been shown that a minor change in input data causes a substantial decrease in model accuracy [14, 15]. To exploit the biometric protection framework, the intruder will use these adversarial examples, resulting in either an unauthorized user having access to the system or an authorized user being unable to access the system.

Szegedy et al. implemented the first adversarial attack, it is called as L-BFGS [16] and it is a costly method of computation. Goodfellow et al. have addressed the shortcomings of the previous system. Another method called FGSM, the Fast Gradient Sign Method, has been implemented, which introduces the degree of disturbance by considering the gradient sign [17]. Goswami et al. suggested an adversarial blackbox attack, which introduces the distortions in the face image and it leads the poor performance face recognition system [18]. The evolutionary algorithm was used by Dong et al. to build adversarial examples [19] and it follows the white box attack strategy. Lu et al. are suggesting FGSM-based attacks, which cause a disruption in all frames of a video. Milton et al have suggested a momentum-based FGSM attack and the CNN model is affected by that attack [20]. Generative Adversarial Networks (GAN) are used in [21] to construct distorted images with regard to samples of face images. In order to create adversarial instances, Rozsa et al. have enhanced the efficacy of the FGSM approach by considering the gradient value, whereas the previous method uses the gradient sign [22]. The DeepFool method has also been used to generate adversarial samples to classify the Lp disturbance that converts the input samples into adversarial data [23].

B. Defensive Mechanism

Two kinds of defensive techniques are used to handle the adversarial attacks, 1. Reactive defensive strategy 2.Proactive defensive strategy [24]. In the reactive defensive mechanism, after the deep learning models are designed, the designer tries to classify the adversarial examples. Whereas the designer aims to build the models more stable until the attacker implements the manipulated samples in the constructive defensive strategy. Few types of proactive defensive methodologies are developing robust classifier, adversarial training, and network distillation. Classifier Models are used as filters to remove the crafted data from the training data which act as preprocessing step. So that the robustness of the model is increased effectively [25].

i) Adversarial example recognition ii) network verification iii) input reconstruction are examples of reactive defensive mechanisms. The binary classifier was considered for the identification of the manipulated samples [26]. The adversarial examples were transformed to approximate original examples in the input reconstruction strategy. In order to recreate the adversarial samples into actual samples by eliminating the perturbations, a denoising auto encoder is used [27]. Network verification, which investigates the input data and tests whether the input violates the characteristics of the deep neural network, is the last technique [28]. In [29], to filter the adversarial instances, the authors used the appropriate dropouts in hidden layers. Agarwal et al. [30, 35, 36] have used the Principal component analysis (PCA) and the Support Vector Machine (SVM) to consider the presence of adversarial attacks.

III. Research Methodology

The Proposed method aims to identify the adversarial examples by removing the perturbations without changing the classifier model. Initially the classifier model is trained with the actual iris images i.e. unperturbed images. In the input examples, the adversarial Iris examples are generated by adding perturbations. To counter this, by using an encoder from a model trained to denoise the perturbations, we aim to eliminate the denoise in the Iris examples. We subsequently decompose the iris example image input into wavelet sub-bands by using wavelet transformation. This defensive mechanism utilizes the convolution layers that are trained to recreate the benign iris images by removing the adversarial noise and it analyzes the mid and high frequencies of wavelet components. In this approach Robust Normalization is used, which has a connection between the removal of outliers in activations and robustness. Dropouts are used to decrease the inter neuron dependencies. Therefore, the neural network is restricted from depending heavily on neuron weights, which could be model vulnerability. To enhance robustness, we suggest using average pooling layers that introduce less loss of information than max pooling layers.

A. Encoder – U-Net Architecture

The goal of this methodology is to extract the perturbations from the manipuated Images and the features of generated Iris examples are retained. For this, a deep convolutional neural network is used which follows an U-net architecture. The U-Net architecture has skip connections that have an effect on problems with gradient vanishing and can transfer image information from convolution layers to deconvolution layers that play a role in reconstructing noisy input. The U-net architecture could learn to denoise and get simple denoised outputs in a stable manner. The explanation why U-net-based denoising models are effective in denoising may be linked to the relations between the contracting path and the expansive path. The U-Net architecture has three sections, a contracting path, a bottleneck and an expanding path. In this architecture, the contracting path utilizes many convolutional operations followed by average-pooling operations. Then the input flows into the expanding layer with corresponding layers of convolution. The contracting and expanding path is linked by the bottom layer. The same is illustrated in Fig. 1.

The necessary preprocessing operations are carried out in the following way: adversarial input image are normalized and reshaped. These images are given as input to the U-net architecture. The encoder layer consists mainly a convolutional layer followed by strong normalization and a dropout layer followed again by a convolutional and robust normalisation layer. Then the corresponding output is applied on Average pooling layer. With the exception of the Convulational Transpose Layer, the decoder portion of the U-net architecture is identical to the encoder part. The representation of the image is fed from the U-net model's earlier layers. That is, from the encoder to the decoder layer. The output is then moved to a convolutional layer to recreate the image without the adversarial noise.

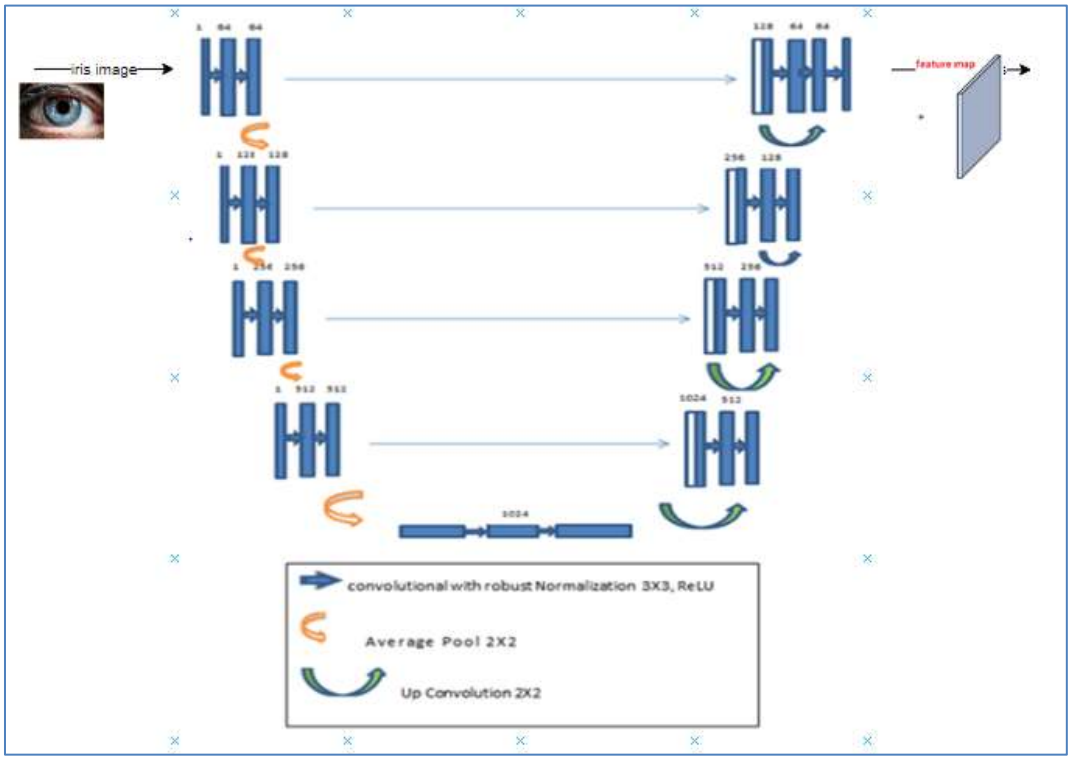


Fig. 1. Architecture of U-Net Denoiser.

The U-net denoising model uses Robust Normalization that outperforms BatchNorm on many datasets for adversarial accuracy while retaining other Normalization advantages. The model is trained to reduce reconstruction errors in order to eliminate the adversarial noise from the adversarial example, so it aims to transform the adversarial examples into their respective benign examples. From equation (1) the reconstruction error is calculated for every batch.

$$reconstruction_{err} = \left|\left|\tilde{x}_i^{(i)} - x_i^{(i)}\right|\right|_2 \quad (1)$$

Where,

$\tilde{x}_i^{(i)}$- reconstructed input

$x_i^{(i)}$- actual input

Without the adversarial noise, the encoder learns the best characteristic representations necessary for reconstruction of the input image. Fig. 2 shows the single instance of the encoder layer.

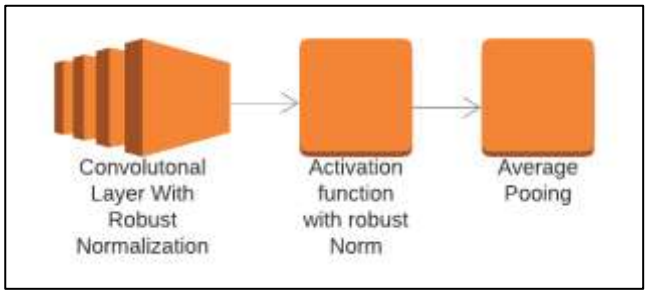


Fig. 2. Single Instance of the Encoder Layer.

B. Wavelet Decomposition

The input image is decomposed in to identical sub bands by using wavelet decomposition technique. This uniform decomposition offers more flexibility for our proposed system to select mid and high-frequency sub-bands. Wavelet image transformation is a very efficient and stable technique and has many benefits. For example, in a digital image, the wavelet analysis preserves the high-frequency edge information and prevents the image from being fuzzy. The method of wavelet analysis is a time-frequency analysis method that selects the appropriate frequency band adaptively based on signal characteristics. In denoising, this property is incredibly helpful as it reduces the loss of data during denoising. In order to achieve optimal reconstruction of the original signal, the wavelet transformation process relies on the best mapping of signals from the actual space to the function space of the wavelet. The proposed solution uses the multi-level discrete wavelet decomposition. This wavelet transformation decomposes the signal into a wavelet range that is mutually orthogonal, and this particular decomposition of the wavelet more finely decomposes sub-bands of low passes. Fig. 3 illustrates the wavelet decomposition stages. The wavelet transform can be expressed by using equation (2).

$$R(a,b) = \int_{-\infty}^{\infty} r(x)\, \phi^*_{(a,b)}(x)dx \quad (2)$$

Where,

* - conjugate symbol

ϕ – Any function, chosen arbitrarily, should follow certain rules

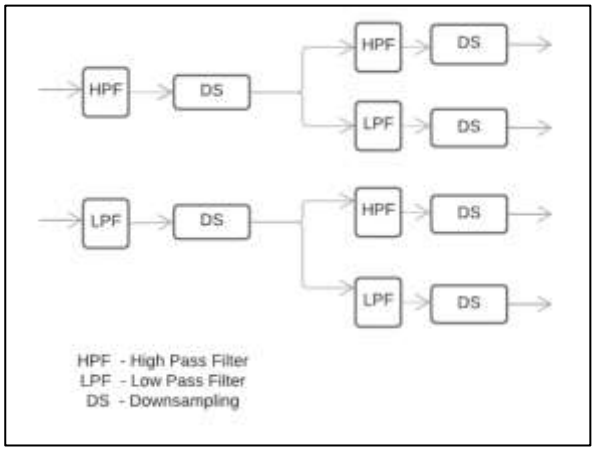


Fig. 3. Wavelet Transform Decomposition Stages.

The encoder's output is fed to the Wavelet layer and the wavelet transform is applied to the image, which splits it into four sub-bands hierarchically. In order to implement the wavelet transforms, we perform a series of operations on each axis to construct partitions. After investigating the directions of low and high pass filters, Multi-level Discrete Wavelet Transformations are determined. For downsampling of the images, even index columns are chosen. The resulting image is then transmitted again to the low pass and high pass filters where the convolved image is generated as an output. The inputs are now down-sampled by rows. This process results in four sub-bands. In these four sub-bands, there are diagonal, horizontal and vertical descriptions of the images.

$$y_{high}[n] = \sum_{i=-\infty}^{\infty} x[i]h[2n-i] \quad (3)$$

Where,

x - input signal

h - high pass filter

$$y_{low}[n] = \sum_{k=-\infty}^{\infty} x[i]l[2n-i] \quad (4)$$

Where,

x - input signal

l - low pass filter

Equation (3) and (4) show the functioning of Low pass and High pass filters with down sampling. All the sub bands are functioning efficiently. The convolution layers which are present in encoder part of the denoising models with Robust normalization are trained to filter the targeted adversarial attack. One more sub band is trained to remove random noise by decreasing the reconstruction error. The deep U-Net architecture is subsequently concatenated by all these sub-bands.

The U-Net Model's output is applied to the convolutional layers with robust normalization, then it has been passed to the global average pooling layer. The purpose of Global average pooling layer is to reduce the number of model parameters drastically and it prevents the overfitting, it results the increase in performance. The average pooling layer of Convolutional Neural Network model doesn't preserve the low level feature sets, but it restores the high level feature map. The results from the previous layers are given into dense layers with dropouts undergoing Robust Normalization. For classification, we have used the softmax activation function. Sparse categorical entropy is the loss metric and Adam is the optimizer.

We integrate regularization in the form of L1 regularizer to avoid overfitting the model. The explanation for preferring L1 rather than L2 is that L1 tends to minimise the coefficients to zero, while L2 reduces them equally. This enables L1 more acceptable for the selection of features, since it helps us to drop any variable with coefficients moving to zero. We observe an increase in the validation accuracy of our classifier by adding L1 regularization.

IV. Results and Discussions

In this part, we discuss the dataset used and how the adversarial perturbations and noise were applied to produce adversarial examples. We conclude the section by describing the findings of the proposed systems and analyzing their efficiency with the previous state-of-the-art mechanisms.

A. Dataset

In the proposed method, by integrating different forms of noise in the clean examples, we produce the adversarial Iris examples. The adversarial dataset is generated by the algorithms FGSM, iGSM and deepfool which are most popular algorithms to produce adversarial examples. Table I gives the descriptions of datasets used in the proposed work. From one model to another, the adversarial noises have remarkable transferability. The perturbations are added in the CASIA Iris V4 Dataset then the Deep CNN U-Net model is trained to remove the noises. The key features required to reconstruct the denoised version of the image from adversarial image are preserved by minimizing the reconstruction error. This can be achieved by using the encoder – decoder layers of the framework.

The wavelet domain decomposition layer belongs to a DCNN denoising model is trained to remove the adversarial noise. The encoder part of this denoiser works as on the wavelet sub-bands. As a whole, using the wavelet transformation function, a single adversarial image is decomposed into four wavelet sub-bands here. Of these four sub-bands, three are trained to remove adversarial noise, while the fourth is trained to eliminate random noise. All these four sub-bands are then concatenated at various layers and eventually transferred into the global average pooling layer. On our final evaluation we observed a rise in the validation accuracy of our classifier. Table II indicates the accuracy of the model before the attack and after the attack. The Deep CNN model is applied on CASIA Iris V4 dataset for classification and the accuracy before FGSM attack is 98.01% whereas after the attack it reduces into 90.24%. The same table indicates the accuracy before and after the attack in case of iGSM, Depfool. The comparison of classification accuracy for the proposed model with existing state of art model is tabulated in Table III.

It is observed that the proposed method is outperformed and the accuracy is good competed to other state of art models. The graphical representation of the comparison is shown in Fig. 4(a) and 4(b).

TABLE I. DESCRIPTION ABOUT THE DATASET

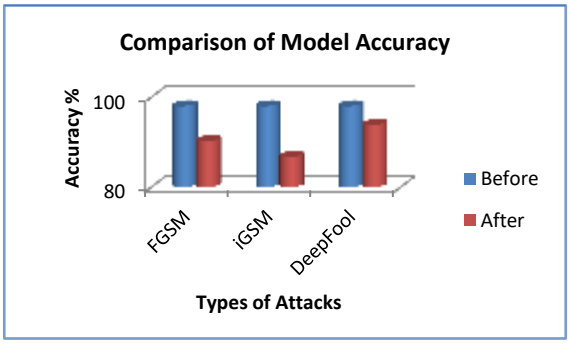
Dataset	Images	Classes
Casia-IrisV4	20000	1000
Casia-IrisV1	1080	108
IITD Database	2240	224
FGSM Database	50000	1000
iGSM Database	20000	1000
Deepfool Database	21080	1000
Noise Dataset	10000	1000

TABLE II. MODEL ACCURACY BEFORE AND AFTER ADVERSARIAL ATTACKS

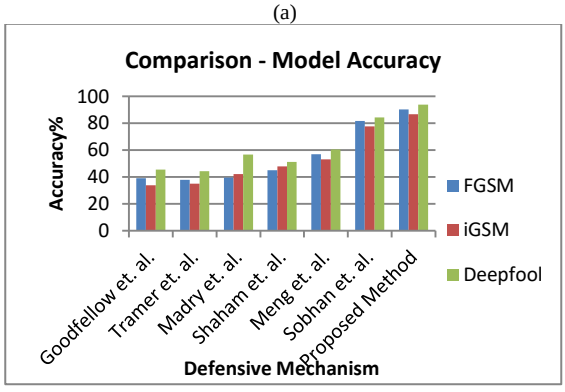
Accuracy Vs Attack	FGSM	iGSM	Deepfool
Before	98.01	98.01	98.01
After	90.24	86.70	93.83

TABLE III. COMPRARISON OF PROPOSED MODEL WITH STATE-OF-THE-ART MODELS IN TERMS OF ACCURACY

Defensive MechanismVs Attacks	FGSM	iGSM	Deepfool
	Accuracy		
Goodfellow et. al. [17]	38.98	33.78	45.47
Tramer et. al. [31]	37.87	34.97	44.41
Madry et. al. [32]	39.51	42.18	56.78
Shaham et. al. [33]	45.15	47.89	51.24
Meng et. al. [27]	57.08	53.26	60.54
Sobhan et. al. [34]	81.65	77.59	84.36
Proposed Method	92.24	86	94.8



(a)



(b)

Fig. 4. (a): Comparison of Model Accuracy before and after the Attack.Three Types of Attacks are Compared- FGSM, iGSM and Deepfool, (b): Model Accuracy for Proposed Methods on Adversarial Attacks.

V. CONCLUSION

Defending adversarial attacks is a crucial move towards reliable implementation of biometrics authentication solutions driven by deep learning. In this proposed work, a novel defending framework has been developed to defend the adversarial attack targeted on Deep Convolutional Neural Networks. Iris recognition system is considered as one of the popular biometric systems which uses the Deep Neural Network for recognition. The proposed strategy is able to detect and reconstruct the adversarial examples consistently. Using an encoder architecture and wavelet decomposition, a framework has been built that takes adversarial input examples and analyzes the wavelet sub bands. Based on the reconstruction error, the framework identifies the attack.

From the Experimental results, it was observed that the proposed strategy was very effective. The proposed framework is compared with other state of art defending strategies and it achieves 92% accuracy during classification of iris images. Further this work can be extended to consider other attack strategies. In this work the wavelet decomposition is applied to detect the adversarial image. In future other equivalent transformation functions like curvelet transform can be applied and study further for other biometric traits.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[2] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, AbdelRahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep Neural Networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82–97, 2012.

[3] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pages 3–14. ACM, 2017.

[4] J. Bruna, C. Szegedy, I. Sutskever, I. Goodfellow, W. Zaremba, R. Fergus, and D. Erhan. Intriguing properties of neural networks. International Conference on Learning Representations, 2014.

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of Neural Networks. arXiv preprint arXiv:1608.04644, 2016.

[6] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In Proceedings of the ACM Asia Conference on Computer and Communications Security, pages 506–519, 2017.

[7] K. Meenakshi and G. Maragatham, "A review on security attacks and protective strategies of machine learning", International Conference on Emerging Current Trends in Computing and Expert Technology, pp. 1076-1087, 2019.

[8] A. K. Jain, K. Nandakumar and A. Ross, "50 years of biometric research: Accomplishments challenges and opportunities", Pattern Recognition Letters, vol. 79, pp. 80-105, 2016.

[9] J. Kim, S. Cho, J. Choi, and R. J. Marks. Iris recognition using wavelet features. Journal of VLSI signal processing systems for signal, image and video technology, 38(2):147– 156, 2004.

[10] S. Soleymani, A. Dabouei, S. M. Iranmanesh, H. Kazemi, J. Dawson, and N. M. Nasrabadi. Prosodic-enhanced siamese convolutional neural networks for cross-device text-independent speaker verification. arXiv preprint arXiv:1808.01026, 2018.

[11] S. Soleymani, A. Torfi, J. Dawson, and N. M. Nasrabadi. Generalized bilinear deep convolutional neural networks for multimodal biometric identification. In 25th IEEE International Conference on Image Processing, pages 763–767, 2018.

[12] F. Taherkhani, N. M. Nasrabadi, and J. Dawson. A deep face identification network enhanced by facial attributes prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 553–560, 2018.

[13] V. Talreja, M. C. Valenti, and N. M. Nasrabadi. Multibiometric secure system based on deep learning. In 2017 IEEE Global conference on signal and information processing (globalSIP), pages 298–302, 2017.

[14] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. International Conference on Learning Representations-Workshop, 2017.

[15] Akshay Agarwal, Akarsha Sehwag, Richa Singh, and Mayank Vatsa. Deceiving face presentation attack detection via image transforms. In IEEE International Conference on Multimedia Big Data, pages 373–382, 2019.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. arXiv preprint, 2013.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[18] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. International Journal of Computer Vision, 127(6-7):719–742, 2019.

[19] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based blackbox adversarial attacks on face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7714–7722, 2019.

[20] Md Ashraful Alam Milton. Evaluation of momentum diverse input iterative fast gradient sign method (M-DI2- FGSM) based attack method on MCS 2018 adversarial attacks on black box face recognition system. arXiv preprint arXiv:1806.08970, 2018.

[21] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. arXiv preprint arXiv:1908.05008, 2019.

[22] A. Rozsa, E. M. Rudd, and T. E. Boult. Adversarial diversity and hard positive generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 25–32, 2016.

[23] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2574–2582, 2016.

[24] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems, 2019.

[25] J. Bradshaw, A. G. d. G. Matthews, and Z. Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. arXiv preprint arXiv:1707.02476, 2017.

[26] Z. Gong, W. Wang, and W.-S. Ku. Adversarial and clean data are not twins. arXiv preprint arXiv:1704.04960, 2017.

[27] D. Meng and H. Chen. Magnet: a two-pronged defense against adversarial examples. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pages 135–147. ACM, 2017.

[28] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In International Conference on Computer Aided Verification, pages 97–117, 2017.

[29] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280, 2017.

[30] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In IEEE International Conference on Biometrics Theory, Applications and Systems, pages 1–7, 2018.

[31] F. Tramer, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204,2017.

[32] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[33] U. Shaham, J. Garritano, Y. Yamada, E. Weinberger,A. Cloninger, X. Cheng, K. Stanton, and Y. Kluger. Defending against adversarial images using basis functions transformations. arXiv preprint arXiv:1803.10840, 2018.

[34] Sobhan Soleymani, Ali Dabouei, Jeremy Dawson, and Nasser M. Nasrabadi, Defending Against Adversarial Iris Examples Using Wavelet Decomposition. arXiv:1908.03176,2019.

[35] Saranya, G., & Pravin, A. A comprehensive study on disease risk predictions in machine learning. International Journal of Electrical and Computer Engineering (IJECE), 10(4), 4217, 2020.

[36] M. K and G. Maragatham, "A Comprehensive survey on Iris Presentation attacks and Detection based on Generative Adversarial Network," 2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2020, pp. 1-9, doi: 10.1109/ICPECTS49113.2020.9336966.

Acquisition of Positional Accuracy with Comparative Analysis of GPS and EGNOS in Urban Constituency

Zeeshan Ali[1]*, Riaz Ahmed Soomro[2], Faisal Ahmed Dahri[3], Muhammad Mujtaba Shaikh[4]
IREA-CNR (National Research Council of Italy), University of Naples "Parthenope", Napoli, Italy[1]
Department of Telecommunication Engineering, Mehran UET, Jamshoro, Pakistan[2, 3]
Department of Telecommunication Engineering, University of Malaga (UMA), Malaga, Spain[4]

***Abstract*—Over the years, precise positioning has been the ultimate goal for Satellite Navigation Systems. The American Global Navigation Satellite System deliver the position and time information intended for various sectors such as vehicle tracking, oil exploration, atmospheric studies, astronomical telescope pointing, airport and harbor security tracking etc. Corresponding technological competitors such as Russian Global Navigation Satellite System (GLONASS), European Union's GALILEO, China's BeiDou and Japanese Quasi Zenith Satellite System (QZSS) are few other versions of Satellite based Augmentation Systems. Nevertheless, stern security measures, geographical statistics and stimulation of diverse Electronic Gadgets at indoor/outdoor surroundings make it critical to acquire data about any vicinity with seamless accessibility, accuracy and integrity with satellite links. In this paper, positional accuracy has been tested with analysis of EGNOS, EDAS and simple GPS receiver models at Rome City, Italy. To support results, various real time experiments/tests has been performed with GPS Receiver SIRF Demo software. The test was conducted on-board a car by installing a laptop equipped with GPS Receiver plus supportive SBAS (EGNOS particularly) through three diverse bus routes of locality and outcomes of few tested samples inside the Rome City center are specified to check the availability of desired satellite signals. Subsequently, comparative analysis has been executed between the simple GPS data received and GPS + EGNOS data collected during daytime traffic. The strength of test signals reveals accuracy of EGNOS in open terrain area with less congestion. Furthermore, Asian and European Advanced GPS systems are compared in terms of performance as well as feasibility of authentic, accurate and swift satellite navigation systems.**

***Keywords*—Differential GPS; augmentation; EGNOS; EDAS; on-board equipment; urban and positional accuracy**

I. INTRODUCTION

In Satellite Communication, Road Vehicle Navigation Systems has emerged potential technology in the domain of Intelligent Transport Systems. Subsequently numerous road applications such as traveler information, automatic emergency calls, route guidance, freight management, advanced driver assistance or electronic fee collection involve On-Board Equipment (OBE) capable of offering highly accurate location obtainable at low cost. To acquire the positional accuracy and integrity with Global Positioning System, EGNOS (European Geostationary Navigation Overlay Service) has been assimilated in European territory; distinctive package as it generates warning messages in case of positional error during satellite navigational calculations. Subsequently, it discards the satellites when false readings appear on display by evaluating certain threshold levels as outcomes [1]. SBAS is designed to grasp satellites navigation signals and broadcast GPS category of signals controlling integrity and wide-area differential correction augmentation data [2]. The EGNOS system superimposes over the GPS and GLONASS schemes to enrich accuracy, availability, reliability and continuity of positional estimation. SBAS concept is better approach in timely correct information of system for integration and correction to the random measurements which leads to the accuracy of coordinates [3]. GPS in urban and mountain areas, where GNSS signals are either blocked or degraded by natural or artificial obstacles, unable to provide accurate positioning due to the poor signal quality [4]. The AUS/NZ SBAS moreover broadcast exact satellite orbits and clock remedies to help ongoing drift equivocalness Precise Point Positioning (PPP) service that can convey 5-20 cm precision [5-8]. The authors have investigated DGPS and EGNOS receivers used vessel maundering in the bay of Gdansk. Two receivers were exploited to record the coordinates [9]. Two measurement sessions were adopted at fixed point to analyze the system performance. In determination of accuracy and integrity of positioning, different approaches were used to compute accurate and precise location. It is noted that variant in calculation of design which meets the integrity requirement of navigation system [10]. In this paper, the availability of effective, seamless and accurate signals has been analyzed by executing few experiments around a local bus route inside the Rome city centre to check the existence of signals during day time traffic. The combination of GPS+EGNOS could be suitable approach to cope up with an accurate positioning problem can improve the accuracy, continuity and integrity of positioning. The outcomes of practical tests confirmed the coverage of GPS and EGNOS signal in urban areas as well as detected the transmission time to a control centre (positioning data acquired from running vehicle). Moreover, the vehicle adopted several bus routes of public transport around the city of Rome.

The rest of the paper is structured as follows: the fundamentals of EGNOS framework is discussed in Section 2. The EDAS architecture and its performance services is explained in Section 3. The experimental setup of the proposed approach is described in Section 4. Section 5 reports the results and finally Section 6 concludes the paper.

*Corresponding Author

II. Fundamental Framework of EGNOS

The broadcasting of integrity messages and differential corrections of satellites has been accomplished with current GNSS setup; integrates by means of network of reference stations positioned across the globe having different names of Satellite Based Augmentation System (SBAS) in diverse constitutions (SBAS intensifies and stabilize the GNSS deployment) [11].

SBAS has been employed by various regions such as Western Europe and North Africa establish EGNOS whereas Wide Area Augmentation System (WAAS) by USA and Multi-Functional Satellite Augmentation System (MSAS) by Japan are corresponding satellite systems [12].

A. EGNOS Structural Architecture

EGNOS, Ranging and Integrity Monitoring Stations (RIMS) accumulate raw GPS data measurements and scrutinize signals quality, multipath mitigation and perceive satellite failure affairs in the processing centers. The user collects corrections and integrity statistics by means of PRN120 and PRN136 as geostationary satellites [13].

The functional skeleton of SBAS sub systems has been indicated in the Fig. 1 that summarizes GPS transmitted data (in some cases GLONASS satellites) to receiving customer. This navigated data has been monitored by networks controlled under SBAS service providers [13].

B. Operation of the EGNOS Segments

EGNOS signal is being broadcasted by the space segment that includes three GEO Space Vehicles (SVs) associated with a unique pseudo random noise; ground control segment manages the system as well as processes EGNOS signals simultaneously. After this activity, diverse categories of users are benefited with the development of EGNOS compliant receivers as the part of user segment. Fig. 2 shows the ground segment of EGNOS. In Europe, 41 Ranging and Integrity Monitoring (RIM) Stations are installed whereas few terminals are deployed at USA and North Africa [14].

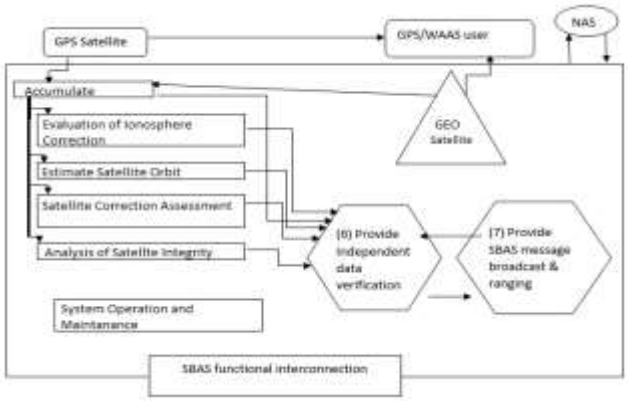


Fig. 1. SBAS Functional Overview.

Fig. 2. EGNOS Ground Segment.

III. EGNOS Data Access Service (EDAS) System Architecture and Performance based Services

EGNOS delivers a terrestrial data service termed as EGNOS Data Access Service (EDAS) with real time data access (EGNOS) is achievable via ground transmission systems. File Transfer Protocol (FTP) offers EGNOS data to authorized customers (e.g., added-value application providers). Subsequently, EGNOS infrastructure (Navigation Land Earth Station and RIMS) acquires data with EDAS single point of access and utilize the real time data to specify performance boundaries [14]. Serially, satellite navigation data engendered by ground stations; EDAS permits users to "plug in" to EGNOS as indicated in Fig. 3 and Fig. 4 represents the EDAS service layout. The acquired data can be utilized efficiently in harsh environment, where signals distracted due to interference or when signals are blocked or invisible [14].

Correspondingly, EGNOS feeds the data to EGNOS Data Server (EDS) and execute number of tasks such as:

- Huge amount of customers is permitted by accepting feasible connection.
- Extra Security Layer has been initiated between customers and EGNOS.
- EGNOS registered data formats and protocols are secured plus EGNOS data has been processed via EDAS services.

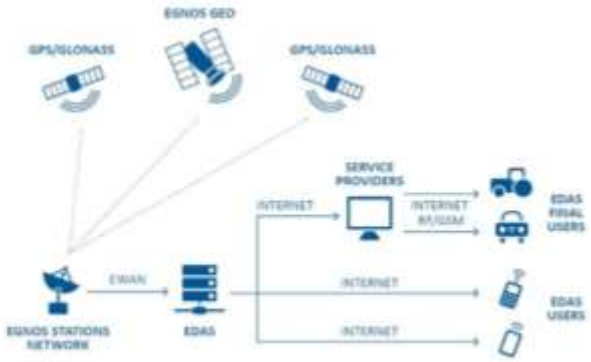


Fig. 3. EDAS High Level Architecture [14].

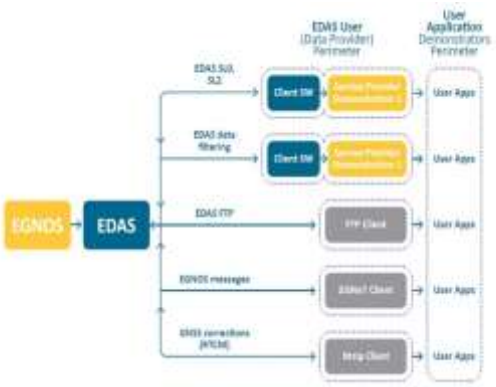


Fig. 4. EDAS Service Flow Diagram [14].

IV. Experiments

A. Test Availability

It has been evident in several FP7/ESA projects that EGNOS is capable enough to uplift the accuracy of GPS receiver in highways (former utilization of EGNOS/GPS receiver was based on transportation of hazardous items). Consequently, EGNOS messages are sent to the receiver via Geostationary Satellite. But, the availability of EGNOS signals via satellite seems to be lacking in the urban areas due to geographical and population dynamics. Therefore, Internet Connection through the EDAS service is an alternate way to deliver EGNOS messages in urban territory.

B. Test Setup

This test is conducted on-board with a GPS receiver connected with the laptop to support SBAS (EGNOS in particular) and the modem linked to the Internet via USB UMTS/HSDPA. To detect coverage, vehicle has pursued the routes of the Rome City Urban Bus Lines in each session as shown in Fig. 4, whereas few test results/samples have been taken away from several regions as mentioned below:

- Around Ponte Principe Amedeo Savoia Aosta, Roma.
- Via Nazionale and Piazza della Repubblica, Roma.
- Corso Vittorio Emanuele II, 00186 Roma, Italy.

After configuring the GPS Receiver Origin ORG1300 to facilitate the SBAS messages, activates the corresponding corrections settlement and yields at least the message GGA (NMEA) positioning with a period equal to 1 second. Along the desired route, the PC has sampled the data generated by the GPS receiver (the manifestation of the GPS /EGNOS positioning and messages) and executed data from adjoining locations and record/trace it with a remote control center via Internet. The successful transaction of data seems to be achievable when network signals exist along the route. The corresponding Fig. 5 indicates the precise real time track during the experiment.

Fig. 5. Accredited Track Pursued Thru the Experiment.

V. Test Analysis of Position Calibration

The Wireless Internet Connection has been utilized to acquire EDAS messages and boosting the accuracy of the localization in urban areas.

A. Performance Comparison between Solitary GPS System and Mutual GPS + EGNOS Signals when Latitude is Not Highly Mounted.

The resultant sample around the area of Ponte Principe Amedeo Savoia Aosta Roma indicates the comparison of solitary GPS and collaborative GPS+EGNOS performance in Fig. 6 and Fig. 7.

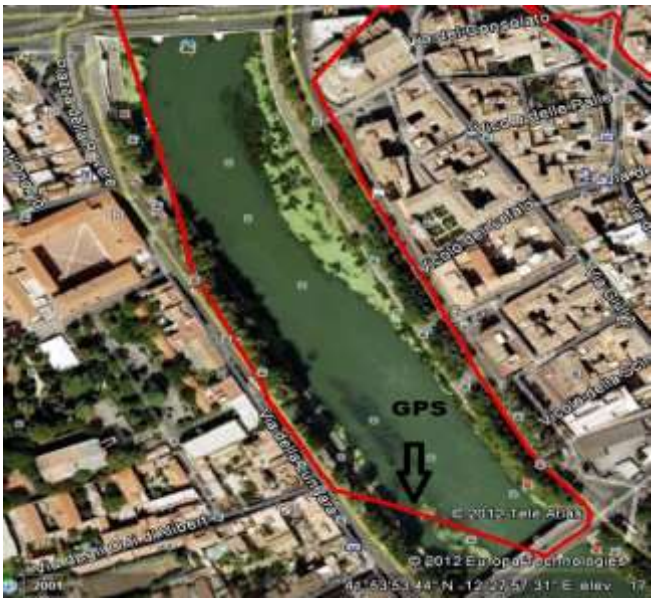


Fig. 6. Solitary GPS Signal Performance.

Fig. 7. GPS + EGNOS Signal.

B. Evaluate the Percentage of Time in which EDAS could Achieve

The experimentation has been analysed which estimates the percentage of available network service and their respective delays such as joint GPS+EGNOS signals and solitary GPS as shown in Fig. 7.

- Blue bar represents the 91.89% availability intervals of EDAS services.
- EDAS and GPS comparative accessibility generates 91.10% (Orange Bar Stats).
- Availability of EGNOS service 42.26% suffered due to inaccuracy in Urban Territory (Specified by Gray Bar).

Though, the availability of EGNOS is quite modest with 42.26% due to urban region obstacles and mega structures as shown in Fig. 5 where accuracy of EGNOS starts to drop and metropolitan canyon disturb precision. However, it harvests astonishing outcomes with improved accuracy when open terrain regions are tested as indicated in Fig. 7.

In the category of satellite navigation augmentation systems, EGNOS enhances the precision of GPS by delivering a positional correctness within three meters. Moreover, GPS receiver without EGNOS measure positional data within 17 meters. Furthermore, the navigation system offers statistics about precision of the location, which is related to the trust developed due to validation of the system's integrity by EGNOS. In addition, it also delivers information about suitable usage of navigation system by activating timely warnings. Integrity feature ensures the accurate tracking of the location in case of emergency situations such as aeronautics and maritime circumstances.

Fig. 8. Availability of Service Representing Scrutinized Networks (In Percentage).

Conversely, EDAS is far more reliable in Urban areas with availability of 91.89% as shown in Fig. 8. Furthermore, EDAS can assist in those spots where the EGNOS and GPS signals are not visible. During tests, it is investigated that attainment of both GPS+EGNOS corrections can detect accurate positions are detected but still GPS+EGNOS signals lags at some locations such as under the tunnels/bridges where there are no visible satellites.

In upcoming future, there are two possibilities for accessibility of best positions – uplift the accuracy of position either by SISNeT or EDAS, especially for those places where the signals are weak or not available. However, SISNeT does not provide safety of life while EDAS is designed in such a way as to provide accuracy as much as the safety of life.

VI. Conclusion

In this paper, the comparative analysis of EGNOS and GPS has been accomplished by accumulating the data on diverse locations of Rome City. It has been witnessed that EGNOS is extremely accurate in timberlands with less congestion in outcomes. However, EGNOS downtown analysis submits inaccuracy in measurements due to hindrances such as tall buildings, bridges, trees and metallic obstructions in the vehicle. With GPS tests, it has been scrutinized that GPS constellation in medium orbits remains unobstructed globally and suits more to urban areas as influence of obstacles seems to be . In open plateaus, though, EGNOS is more precise in contrast to GPS with available data positions. Currently, Pakistan is substantially lacking its own Augmentation System and relying on the American GPS Systems (global accuracy of 10-15 meters) to stipulate statistics about Route Guidance in Pakistan Territory. Though, space agency of Pakistan- SUPARCO have joined their hands with P.R. China to launch their own augmentation system in the upcoming days.

ACKNOWLEDGMENT

This research work has been carried out in the Rome City Center by ensuing number of bus routes and steered on-board a car. We would like to thank Prof. Dr. Ernestina Cianca for her indispensable guidance and assistance throughout the project as without her ideas stream of research tasks were not attainable. The author would also like to thank Dr. Riaz Ahmed Soomro for the guidance.

REFERENCES

[1] European Global Navigation Satellite Systems Agency (GSA), “EGNOS Service Definition Document, Safety of Life Service (SoL) Issue 3.0, ISBN: 978-92-9206-025-1 can be accessed via: https://egnosportal.gsa.europa.eu/sites/default/ files/uploads/Brochure_SoL2015_150924_High Def.pdf.

[2] Kaplan, Elliott D. and Hegarty, Christopher, "Understanding GPS: Principles and Applications", Second Edition, Artech House Publishers, 2005.

[3] Felski, Andrzej, Aleksander Nowak, and Tomasz Woźniak. "Accuracy and availability of EGNOS-results of observations." Artificial Satellites 46, no. 3 (2011): 111-118.

[4] Angrisano, A., S. Gaglione, and C. Gioia. "Performance assessment of GPS/GLONASS single point positioning in an urban environment."Acta Geodaetica et Geophysica 48, no. 2 (2013): 149-161.

[5] Zumberge, J. F., M. B. Heflin, D. C. Jefferson, M. M. Watkins, and F. H. Webb. "Precise point positioning for the efficient and robust analysis of GPS data from large networks." Journal of geophysical research: solid earth 102, no. B3 (1997): 5005-5017.

[6] Héroux, P., Y. Gao, J. Kouba, F. Lahaye, Y. Mireault, P. Collins, K. Macleod, P. Tétreault, and K. Chen. "Products and applications for Precise Point Positioning-Moving towards real-time." In Proceedings of the 17th international technical meeting of the satellite division of The Institute of Navigation (ION GNSS 2004), pp. 1832-1843. 2004.

[7] El-Mowafy, Ahmed, Manoj Deo, and Nobuaki Kubo. "Maintaining real-time precise point positioning during outages of orbit and clock corrections." GPS solutions 21, no. 3 (2017): 937-947.

[8] El-Mowafy, Ahmed, and Nobuaki Kubo. "Integrity monitoring for Positioning of intelligent transport systems using integrated RTK-GNSS, IMU and vehicle odometer." IET Intelligent Transport Systems 12, no. 8 (2018): 901-908.

[9] Specht, Cezary, Jan Pawelski, Leszek Smolarek, Mariusz Specht, and Pawel Dabrowski. "Assessment of the positioning accuracy of DGPS and EGNOS systems in the Bay of Gdansk using maritime dynamic measurements." The Journal of Navigation 72, no. 3 (2019): 575-587.

[10] Grunwald, Grzegorz, Mieczysław Bakuła, Adam Ciećko, and Rafał Kaźmierczak. "Examination of GPS/EGNOS integrity in north-eastern Poland." IET Radar, Sonar & Navigation 10, no. 1 (2016): 114-121.

[11] Augmentation and integrity monitoring network and EGNOS Performance Comparison for Train Positioning. IEEE Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European. Date of Conference: 1-5 Sept. 2014. Electronic ISSN: 2076-1465.

[12] Simo Marila; Mohammad Zahidul H. Bhuiyan; Jaakko Kuokkanen; Hannu Koivula; Heidi Kuusniemi.- Performance Comparison of Differential GNSS, EGNOS and SDCM in Different User Scenarios in Finland IEEE 2016 European Navigation Conference (ENC).

[13] EGNOS Open Service (OS) Service Definition Document, EGNOS Open Service (OS) SDD, Issue 2.1. Accessed via: https://egnos-portal.gsa.europa.eu/sites/default/files/EGNOS_OS_SDD_2.1. Pdf

[14] Egnos Data Access Service (EDAS) Service Definition Document, SDD, Issue 2.0, April 2013, EGN-SDD EDAS, V2.0. Accessed via: http://www.galileoservices.org/news_events/edas_sdd_v2_0.pdf.

Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies

Nikola Modrušan[1], Leo Mršić[3]
Algebra University College
Zagreb, Croatia

Kornelije Rabuzin[2]
Faculty of Organization and Informatics
Varaždin, Croatia

***Abstract*—Numerous studies and various methods have been used to detect and prevent corruption in public procurement. With the development of IT technology and thus the digitization of the Public Procurement Process (PPP), the amount of available data is increasing. Studies have shown progress in this area and have revealed many challenges and open issues geared to the various goals outlined in this paper. Different data mining and business intelligence techniques and methods are being used to develop models that will find any suspicious public procurement process, contracts, economic operators, or to classify observations as corrupt. In addition to using classification models, methods such as association rules and graph databases are used to find relationships between economic operators and contracting authorities, as well as to find daughter companies that participate in PPP collusion. Therefore, this paper addresses a comprehensive review of the emerging techniques and models used for the detection of suspicious or corrupted observations, their goals, open issues, challenges, methods and metrics used, tools, and relevant data sources. The findings show that models are mostly fitted on historical data and move in the direction of an early warning system. Moreover, the efficiency of fraud or anomaly detection depends on data set quality and detection of the most important red flags. The study is presenting a summary of identified fraud detection model objectives such as predicting fraud risk in contracts and contractors or finding split purchases, and detection of used data sources such as public procurement process or economic operator data.**

***Keywords*—Public procurement; fraud detection techniques; corruption detection; fraud detection review; fraud data source**

I. Introduction

Public procurement is a process through which the state orders different types of commodity services and thus spends public money. Accordingly, the public always raises questions about lawful spending and whether the public sector is getting the best service or goods for a real price or if there is some form of corruption that generates a loss of their money [1]. Corruption in public procurement is defined as the abuse of power for private profit [2].

Public procurement integrity is a term often used in the literature and is defined as the use of funds, resources, assets, and authority, according to the intended official purpose, to be used in accordance with the public interest [3]. All acts that are not under this definition can be considered a violation of integrity, and therefore they can be proclaimed as suspicious or criminal behavior. Such acts may occur at different stages of the public procurement process, from the creation of tender to the implementation, documentation, contract making, and realization [4]. The most common types of procurement fraud and corruption are bid-rigging, collusion between vendors and employees, and collusion between vendors [5]. Table I shows that there exist a lot of different fraud and corruption types and the most interesting area is certainly finance or accounting and the public sector. For each type, different red flags and corruption indicators that are specific and represent a correlation with corruptive actions are detected [4,6]. Moreover, Table II shows types of corruption, information about the impact of each corruption type, and level of occurrence probability. This result presents a good starting point in dealing with corruption and the fact that bribery and kickbacks, conflict of interest, collusive bidding, implementation, donations to political parties have the highest fraud impact.

TABLE I. Most Disruptive Fraud Events by Industry- Adopted According to [10]

Rank	Energy, Utilities, Resources	Financial Services	Gov/ Public Sector	Health Industries
1	Bribery and Corruption 17%	Customer Fraud 27%	Cyber-crime 17%	Cyber-crime 16%
2	Asset Misappropriation 16%	Cyber-crime 15%	Financial Statement Fraud 17%	Financial Statement Fraud 13%
3	Financial Statement Fraud 13%	Financial Statement Fraud 14%	Bribery and Corruption 16%	Customer Fraud 13%

TABLE II. Probability and Impact of Corruption Risks- Adopted According to [4]

Type of corruption	Impact	Probability
Bribery and kickbacks	High	Medium
Conflict of interest	High	Medium
Collusive bidding	High	High
Shell companies	Medium	Medium
Leaking bid data	Low	Medium
Unbalanced bidding	Low	Medium
Manipulation of the bidding procedure	Low	Low
Split purchases	Medium	Low
Rigged specifications	Medium	Medium
Excluding qualified bidders	Medium	High
Unnecessary purchases	Low	Medium
Implementation	High	Medium
Donations to political parties	High	High

On average, corruption accounts for 5% of the total value of public procurement, which is around 14% of the European Union's (EU) GDP, or EUR 1.9 trillion within the EU, which is one of the main reasons why in former years many efforts have been invested in the field of corruption definition and detecting suspicious actions [5,7]. A 2020 study by the Association of Certified Fraud Examiners published the Report on Professional Fraud and Abuse. This report provided the results of an analysis of 2,504 cases of professional fraud that occurred in 125 countries worldwide [8,9].

Regarding studies and the fight against corruption in this segment, country authorities use various techniques mostly focused on regulating the public procurement process by using different questionnaires and establishing process control; however, the conclusions of the study clearly state that a correctly set public procurement law is insufficient, and there is a lack of control mechanisms for prevention [5]. Fraud committed by those you invited in (e.g., internal perpetrators, vendors/suppliers) represent nearly half of all fraud reported [10]. By Table III in the process of making public procurement fraud, the most responsible is the middle management– more than 37% of reported cases.

TABLE III. WHO'S COMMITTING FRAUD - PERPETRATORS: EXTERNAL, INTERNAL AND COLLUSION BETWEEN THEM - ADOPTED ACCORDING TO [10]

Perpetrator	Reported	Top perpetrator
External preparator	39%	1. Customer 26% 2. Hackers 24% 3. Supplier 19%
Internal preparator	37%	1. Middle mgmt. 34% 2. Operations staff 31% 3. Senior mgmt. 26%
Collusion between internal and external	20%	-

There are several types of methods for corruption detection and measurement: surveys, administrative data from crime statistics, ombudsmen, pp offices, supreme audit institutions, pp governance risk assessments, and analyses of contracts [11]. A World Bank study presents a few major technology trends for public sector fraud and corruption such as big data, cloud computing platforms, artificial intelligence, and machine learning, biometrics (ID4D), FinTech digital money, distributed ledger technology or blockchain, and the Internet of Things (IoT) [8].

Nevertheless, it is difficult to create efficient corruption detection models if there isn't enough quality and diverse data. So, it is widely accepted that access to public information increases the level of transparency in the fight against corruption [11].

Each country has self-organized state-level preventative and anti-corruption agencies responsible to establish the mentioned procedures and monitor law enforcement. With the somewhat onward digitization of the public procurement process, there is an ever-increasing amount of data that is unconnected and largely unstructured; but, with some effort and specific techniques, scientists can use that data to analyze the public procurement process and find adequate corruption indicators. In this study, the approach of detection of public procurement corruption using advanced digital techniques and data models will be explored. With this modus, the study entered the Big Data area, where various advanced statistics and data mining techniques are used to elicit such knowledge. Thus, the fight against corruption in the public procurement segment is not a novelty.

Previous research related to the literature overview of using emerging techniques (e.g., Artificial intelligence) in public procurement fraud detection has four research questions: what are the characteristics of the organizations in which the investigations are carried out, the technological tools, and data mining methodologies and techniques [42]. The focus of the mentioned detection methods is based on data from public procurement contracts. Detection methods, as well as techniques, largely depend on the input data set so one goal of the study is to find and summarize the data sets and methods used for the detection of fraud in the public procurement process. Besides tools and methods, their metrics, challenges, and open issues, the relevant question is what indicators or red flags are used. Most emerging advanced technologies depend on data labeling, not only detecting corruption but also anomalies and suspicious tenders. The open question is how corruption is defined because models are estimating the probability of corruption, predicting the number of bidding tenders, predicting fraud risk in contracts and contractors, finding split purchases, etc. In this research, systems that use advanced technologies and tools for detecting anomalies, fraud, and suspicious public procurement procedures, although they represent modules and closed systems about which there is not enough public information will be detected. Overall, the study will try to obtain more robust results.

The paper is organized as follows: Section 2 presents an overview of the research done in the field of detecting fraud in public procurement by using data mining techniques and machine learning models. The section is divided into subsections where conclusions about the models, methods, metrics, data labeling approach, and corruption detection indicators used are presented. Section 3 represents a short description and list of tools that are used for analysis, monitoring, or fraud detection in the area of public procurement. In Section 4, open issues and further research opportunities are highlighted. Finally, conclusions in Section 5 are provided.

II. PUBLIC PROCUREMENT FRAUD DETECTION

By analyzing scientific databases (SCOPUS, ScienceDirect, Google Scholar, and Web of Science) in the period of last 5 years, after segmentation, a total of 23 scientific studies that are relevant to the study area have been detected and reviewed to gain this literature overview. The main inquiry is made from the combination of next keywords: public procurement; public procurement fraud, public procurement anomalies, public procurement indicators, public procurement red flags, public procurement application, public procurement system, procurement data mining, public procurement methods, public procurement artificial intelligence). Certain studies were focused on legal and organization frameworks, interviews, or statistical models so they are excluded from this overview. Within the scope of

public procurement, different procedure types were found, such as open, restricted, and negotiated procedures, auctions, etc., for which different public procurement rules apply. Different rules result in different processes, and with a lot of dissimilar corruption indicators, this complicates fraud detection [4]. Accordingly, [12] dealt with collusion detection in auctions and provided a review of the methods and data set characteristics. The authors concluded that a large amount of different data is needed for the purpose of quality model results. Even though they described their research goals, the lack of described techniques, models, and the data mining process was noticed in this review. On the other hand, [13] of the total of six studies cited in the literature review, three studies related to the detection of procurement corruption. In contrast, others were related to segments such as the supply chain and the economic sector, and thus they are not relevant to this study. How approaches intersect and created a complex matrix that can be structured using technology and AI support were identified. Diversity in the approaches used in the selected cases leads to the main question: "What were the goals of the studies and which methods are used for fulfilling them?" Therefore goals of the studies and models with the used methods were extracted (Table IV). Also, for this inquiry, two more pieces of information are interesting. Therefore, dealing with the classification of observations or corruption prediction, it is interesting to find what kind of features or data are used to proclaim some observations as bad, suspicious, or corrupted and what metrics are used in order to compare results.

A. Corruption Detection Methods and Models

The use of various analytical and statistical methods was discussed by [5,14]. According to them, corruption detection was first done in the telephone, insurance, and banking industries, which takes a lot of time and domain knowledge from various areas including, legal, financial, commercial, and others. By data in Table IV, it is important to emphasize that the research in this segment is largely focused on the development and application of predictive models and the detection of relationships between economic operators and contracting authorities. In essence, this is a complex matter and is composed of statistical methods, various data mining methods, and machine learning. The literature review shows that researchers used two very familiar approaches, namely supervised and unsupervised learning. These methods differ in target variables, that is, in supervised learning, we have precisely defined target variables as the output of the model, while in unsupervised learning we do not have pre-set variables; so, the models are suitable for seeking anomalies. Still, depending on the model, improvement sometimes is needed to add classified observations [15]. It is important to emphasize that such models are used to detect anomalies, which may be the subject of analysis in some later steps [16]. In general, almost all studies show that the fraud detection model is divided into few steps showed in Fig. 1.

The most commonly used methods in the studies are linear and logistic regression, neural networks, and Naive Bayes algorithms since they are most used for classification and clustering. Namely, models are fitted on historical data and move in the direction of an early warning system that can provide pre-determined supervisory bodies with insights into the risks associated with concluding contracts with risky economic operators [1,18,21] or can identify potential cartels or collusion behavior using associative rules or graph databases algorithms to see the relationships between economic operators and eventually their daughter companies [12,25,26,29,32,43].

The observed studies and created models are used for several different purposes in the detection of corruption in public procurement and at various stages of the public procurement process. Following the observed studies by Table IV, the summary of identified objectives is:

- Estimating the probability of corruption
- Predicting the number of bidding tenders
- Predicting fraud risk in contracts and contractors
- Finding split purchases
- Anomaly detection
- Regression analysis to predict more sensitive features of a procurement
- Detecting anomalies
- Cartel detection
- Collusive behavior
- Conflicts of interest
- Detection of fraudulent public procurement processes

B. Corruption Indicators

One of the essential segments and research questions is certainly the input data. The studies are focused on the detection and analysis of high-quality corruption indicators, risk patterns, or red flags as representatives of corruptive or suspicious actions with the aim of developing models with the best predictive features [3,4,5,6,22,24,31,32]. Including different databases, pattern recognition, and elicitation knowledge is part of the Knowledge Discovery area. In short, studies have suggested that by applying the Big Data approach and data mining methods, better results will be achieved, and better indicators can be found [6,32].

For this very reason, sets of input data are being attempted to expand with different kinds of databases (Fig. 2) to create a data lake or unified data set that can support patterns of suspicious or even corrupt behavior in the procurement process. Certainly, the quality of fraud or anomaly detection depends on the quality of the red flags.

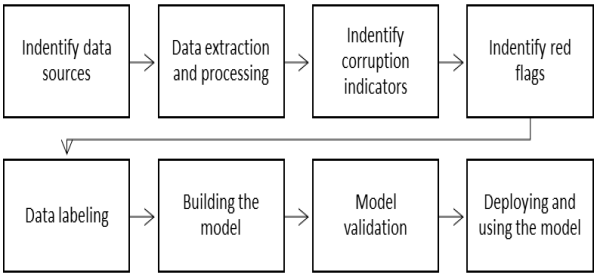


Fig. 1. Fraud Detection Model.

TABLE IV. LITERATURE REVIEW

Paper	Goal	Model/Methods	Target variables
5, 6	Estimation the probability of corruption	Probit - Linear Regression	Corrupt, Clean cases
17	Predict the number of tenders	k-NN, LibSVM, LibLinear Ensemble, Neural network	"Suspicious ": Single bid tenders
18	Predictive model of fraud risk in contracts	One-Class Support Vector Machine, Logistic Regression	"Risky ": Excluded contractors because of fraud, corruption, violation of anti-trust laws
1	Prediction of malfeasance within contracts	Lasso Logistic Regression, Conditional Inference tree, Gradient Boosting machine	"Suspicious ": Extensions to contracts, sanctioned contractors, blacklist contractors
19	Split purchases	Tree Augmented Network, Bayesian Networks	"Suspicious ": Same institutions on the same month and year that added up to more than 8,000E
20	A predictive model of fraud risk in contracts	Naive Bayes, Tree-Augmented Naive Bayes score-based learning algorithms	"Risky ": Temporary suspension of the bid, declaration of non-trustworthiness, impediment to bid and hire.
21	A predictive model of fraud risk in contracts	Logistic Regression, Decision Tree	"Risky ": Supplier serious errors in the execution of any contract
22	Prediction models of public procurement irregularities designed for initial screening of contractors	A neural network, Deep Neural Network, Logistic Regression, Discriminant Function Analysis	"Risky ": Bidding company receives at least one severe penalty due to the serious irregularity
23,24	Coefficients that represent the strength of association between each underlying likely corruption input and likely corruption outcome	Logistic Regression, Linear Regression	"Suspicious ": Winner's Share of Issuer's Contracts, Single Bidder, Exclusion of All but One Bidder
25	Cartel detection	Clustering, association rules, multi-agent approach	Relationship between companies
13	Cartel detection	Association rules – A-priori algorithm	Relationship between companies
16	Anomaly detection	Deep Learning Auto-encoder algorithm	Anomaly
12	Uncovering the structure of collusive behavior	The reduced form of linear regression enriched KRLS method with the CF approach	Relationship between companies
26,43	Identify relationships between companies	Graph databases, decision support system, rule-based	Entities involved in the process
27	Detection of fraud public procurement processes	Naive Bayes, Bayesian networks, decision tree, and neural network.	"Suspicious ": Court rulings, Komisi Pemberantasan Korupsi (KPK) publication, and public comment
28, 29, 30, 31	Detection of suspicious public procurement processes	Data mining, linear regression, Support vector machines, Naive Bayes, Process mining	One bid tender - single bid
32	Collusion between bidders, conflicts of interest, and companies owned by a potentially straw person used for disguising its real owner	Graph theory, clustering, and regression analysis with advanced data science methods	Collusion risk patterns, Company-level risk patterns, Person-level risk patterns

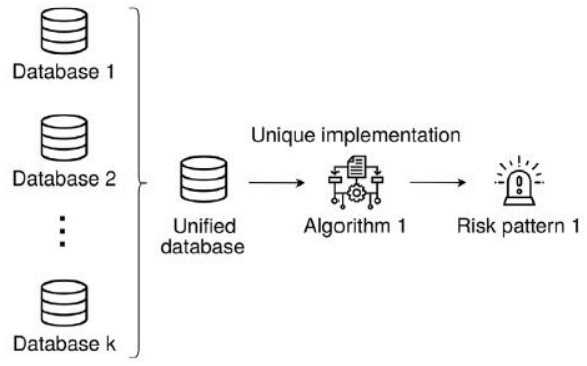


Fig. 2. Scalable Data Unification: the Algorithm that obtains each Risk Pattern is Implemented only once - Adopted According to [32].

More than 200 different indicators are known so far and are used as variables in algorithms, models, or techniques to perform some of the identified objectives in Table IV [31,32,33]. Due to space limitations, a few of them will be mentioned, as follows:

- Unusually short deadline between the announcement of the tender and the deadline for submission of bids.
- The time between the announcement of the tender and the signing of the contract.
- A high percentage of administratively rejected bids in the procedure.
- An unusually small number of correct bids at the level of the procurement procedure.

- A bid accepted before the deadline for submission of bids.
- High ratio of the value of contracts signed under special conditions in relation to the total value of all contracts of an individual client, etc.

In the same way, selecting non-open and less transparent tender procedures reduces the number of possible bids and opens space for awarding a contract to the same well-connected company [23].

Fazekas and Toth used linear regression to find the most useful indicators. Nevertheless, there exist a lot of white papers or studies that present corruption indicators [4,31]. Still, the problem is always choosing and using the right indicators even if we have an indicator that doesn't mean that we have a right and useful red flag. The process of getting indicators seems to be manual by using the expert's domain knowledge, interviews, or surveys [3,4,28]. Authors have searched for different kinds of methods to automate and improve this process. They have implemented dimensional reduction to reduce and include indicators with the best performance using Correlation Analysis (CA), Principal Component Analysis (PCA), and Weighted Principal Component Analysis (WPCA) [27].

Data collected by government bodies or agencies are attempted to be merged, meaning data on the contracting authority or the economic operator, the people who run the company, political connections, etc. These data are actually the attributes needed for the model to make a conclusion or an output prediction, and if we are in a large area of input data, this data needs to be normalized. For this purpose, the Big Data approach is used to process the data in various ways and format it in a model-suitable format, e.g., text-mining techniques such as word tokenization, vectorization, and steeming are used in word processing [17,29]. It is also important to note the application of the above-mentioned method to documents that are a major part of the tender [28]. Keeping all this in mind, it is important to extract the knowledge from a set of data and find patterns and correlations between variables. From the results in Table IV, the used data sets with a few examples can be summarized as:

- Public procurement process data (e.g., type of procedure, estimation price, data type attributes, number of bidders, call for tenders' modification, process duration, tender documentation).
- Economic operator data (e.g., board members, address, contact person, annual tender plan).
- Contracting authority data (e.g., owners, daughter companies, partners, address, telephone).
- Contract data (e.g., price, contract extension, duration date).
- Electronic invoices with products data (e.g., unit of measure, a specific product, product quantity, product price);
- Databases of sanctioned contractors; blacklist contractors; court judgments (corrupt cases); political ties.
- Banking records containing specific details of each transaction.

C. Data Labeling

The next significant observed segment is the attributes, according to which certain models learn to recognize or detect certain prediction classes (mentioned target variables). It has been noted in the studies by [5,6] that only a small number of authors have a clearly specified data set that contains information on whether competition was corrupt, which would mean that there must be a verdict regarding a particular procurement process or a valid classification from that of a superior's institutions, which is not the case in all countries. For this purpose, the authors have taken different features to make some observations suspect, bad, or risky (not necessarily those names) and thus have created prediction classes and introduced certain metrics for that segment, e.g., "Suspicious": single bid tenders, extensions to contracts, sanctioned contractors, blacklisted contractors, and the same institutions on the same month and year that added up to more than 8,000€; e.g., "Risky": excluded contractors because of fraud, corruption, violation of anti-trust laws, temporary suspension of the bid, declaration of non-trustworthiness, impediment to bid and hire, suppliers' serious errors in the execution of any contract, and bidding company receives at least one severe penalty due to the serious irregularity.

As part of the current analysis, it is noted that the investigations aimed at establishing models of detection of corruption risks related to the execution of contracts or corruption by the Economic Operator are based on data contained in databases where irregularities in the execution of contracts have been reported due to fraud, corruption, or violation of anti-trust laws. On the other hand, calculation of corruption risk or classification of corrupt PPP is based on a variable such as the "number of bids," where the aim is to predict if the tender will end up with one bid [17,23,24,28,29,30,31]. The authors have proclaimed these kinds of observations as suspicious.

D. Metrics and Results

Objective testing evaluation requires appropriate methods for accurate measurement. The most commonly used measurements are accuracy, recall, and precision based on a confusion matrix that contains data about the number of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) results. Thus, accuracy is about the proportion of exactly graded examples in the set of all examples. Precision tells us what part of precisely classified examples is in a set of positively classified examples and recalls the part of precisely classified examples in the set of all positive examples. These measures may, in some way, describe a model, but in order to find out the true power of the model, other measures that represent the relation between them should be used. The ROC (Receiver Operating Characteristic) curve is a graphical representation of the binary classifier performance and the area under the ROC curve is called AUC, as it provides a general

evaluation of the model and suggests the ability of the model to discern between the two classes [20,34]. Another very used useful metric in the case of linear models is r2 (pronounced "R squared"), which ""measures the proportion of variation in the responses explained by the available predictor" [35].

Although there are no patterns used, the metrics themselves are different. Namely, accuracy ranges between 30% [16] and even 99% [24], which is pretty "bad", but also are results that are too good that even the authors commented on it. A similar situation is with recall and precision moving at similar intervals. What needs to be emphasized is that, in some studies by [4,26,34], other than accuracy, no other relevant metric is mentioned, which is not enough. AOC metric told us about the power of the model and was about 0.87 for the study that had abnormally high accuracy (more than 99%). [19] R2 was used by Fazekas, as its purpose is to show the power of the linear model. In some studies, r2 varies between 0.2 and 0.55. Detection of cartels and collusion behavior is based on associative rules or graph databases algorithms where the aim is to find relationships between economic operators and contracting authorities. The basic measure in this area is an indication of how often the rule has been found to be true named confidence. Process mining has also been used to analyze the differences between single- and multiple-bid tenders. Process mining has proved that procedures with more than one bid do last longer and that some single bid tenders lasted an extremely short period of time [31].

III. Public Procurement Tools

In previous chapters, some conclusions about the topic of the used methods, input data, labeling data, and metrics were made. All these components ultimately met conditions to create a system for monitoring or analyzing PPP. It is clear that the availability and reliability of the data are the basic premise for the model and can produce the best results. It is necessary to have quality and reliable communication between different state bodies and the connection of all relevant data that the model can use. Otherwise, the lack of the mentioned components can make the process of monitoring and the analysis of all these data quite complex [36].

Even though public procurement processes are defined by policy acts, states independently develop and digitize their systems. Croatia implemented a Public Procurement System (PPS) named EOJN (https://eojn.nn.hr/), which is fully electronic, but without any analytics or monitoring tools. India (https://eprocure.gov.in/) has the eProcurement System, which enables the Tenderers to download the Tender Schedule free of cost and then submit the bids online through this portal. The Irish government implemented an electronic tendering platform administered by the Office of Government Procurement (https://www.etenders.gov.ie/). "The site is designed to be a central facility for all public sector contracting authorities to advertise procurement opportunities and award notices". Some of the countries developed one central platform for PPP, but some have more than one managed by private sectors, for example, Germany or Austria. In the case of multiple PPS, there is a need to have one portal where economic operators can have an overall view of all tenders. For example, the USA have a private project named Tendersinfo (https://www.tendersinfo.com/) as an "online government Tender information provider company, helping business across the globe in finding business opportunities". On the EU level, there is TED (Tenders Electronic Daily https://ted.europa.eu) as an online version dedicated to the European public procurement overview with an amount of 746 thousand published procurement award notices yearly, including 235 thousand calls for tenders worth approximately €545 billion.

The basic components are certainly electronic forms for bid submission, but part of the public procurement system also contains tools for analysis and monitoring of the entire process, whose main purpose is to generate reports, monitor budget spending, and research anomalies. The European Commission and the member states protect their financial interests by using advanced technologies and by the digitalization of the public procurement process itself. Of course, as part of such processes, it is necessary to change and adjust the laws and policies that result from it [3,5,37]. As part of the research, several advanced systems in the field of public procurement were detected and will be described in the continuation of this paper.

Brazil's decision support system for fraud detection in public procurement is a robust tool implemented with the aim of systematic analysis and the identification of the main risk patterns, such as collusion between bidders, conflicts of interest, and risk companies using algorithms such as graph theory, clustering, and regression analysis with advanced data science methods [32]. A similar tool was developed in Africa, named Tendersure (https://www.tendersure.co.ke/), which is based on web technology but does not use advanced technologies and tools such as the system from Brazil. In Ukraine, as part of the national public procurement system, there is the DoZorro tool, which is based on artificial intelligence or supervised learning, and its purpose is to find suspicious tenders depending on risk indicators [33]. The Red Flag system in Hungary (www.redflags.eu) was created in a similar way. Its purpose is to detect risky public procurement procedures and thus present an early warning system. The system is still at an early stage of development. As can be seen from the details of the tools, not all tools are based on advanced algorithms or some form of artificial intelligence; some are also analytical and statistical tools. In Croatia, there is the Integrity Observer System (http://integrityobservers.eu), which is in the form of a dashboard based on data collected directly from the electronic public procurement system and data collected from interviews with the local community. The system is like ERAR (https://erar.si/), which is an online service made in Slovenia. That service provides information on the flow of public money and is linked to contracts between economic operators and the contracting authority. To give the public efficient and transparent public procurement procedure analytics, each country has its own electronically public procurement system that has at least some of the application modules adapted for such purpose (publicly published contract register, register of procurement plans, payment records, etc.). For international transparency, some of them are Macedonia, Georgia, Slovakia, Poland, etc. [38].

It is important to emphasize that when the public procurement process is subject, tools designed to analyze the distribution and use of public money of European funds were included, whose contracting processes must also be carried out by public procurement laws. The European Commission has implemented a risk assessment tool, ARACHNE, to detect and prevent projects that are vulnerable to fraud, conflict of interest, and irregularities [5,6,20]. In addition to the mentioned tool, an analytical database with all the information about users and projects funded by the Directorate-General Communications Networks, Content, and Technology was created with the aim of detecting links about people and projects with all their data, such as phone numbers and addresses. This system is not a warning system, but it is used in cases of doubt if there are irregularities in the project [5]. DAISY is a tool for data mining developed by the Directorate-General for Research and Innovation with the purpose to identify links between users of funds in the scope of research projects. DAISY is used when there is a suspicious fraud of the specific user of the funds [5].

A lot of tools to mitigate fraud (Table IV and Section 3) or public procurement corruption detection was detected. It is difficult to find information on how they work and what methods of corruption detection uses. This is one of the open questions and action points for further research.

IV. Future Research

There are several open issues related to the topic of estimation and detection of corruption in public procurement, which the authors have mentioned in their studies. Therfore, the red flags or corruption indicators are some of the most important points in the detection of fraud, since the segment is heavily dependent on the prediction itself. Fraud detection isn't a novelty; it is widely used in different areas, such as banking, insurance, company procurement, etc. [40]. The authors state that further work is needed to investigate the ranks of red flags and filter them in a certain way, as well as the interaction between institutions that monitor corruption in public procurement, all with the aim of a more precise corruption estimate [17,23].

To find the most important red flags, different methods are used. The main part of this activity is just a manual job, so to significantly improve the process and make it automated, the authors propose using entity recognition techniques [27]. The aim of the model is, in most cases, to get the best precision. In one study [19], the authors obtained almost ideal results, i.e., metrics around 0.99, and concluded that further research is needed to understand why the results are so good that an analysis can result in some discoveries in the relationship between variables. Also, further analysis is proposed, but in the segment of different types of fraud. Thus, the idea is to include new indicators that will cover the new cases previously ignored as well as the use of optimized algorithms in the parameterization of models [20].

Although the use of advanced data analysis techniques and knowledge elicitation was already identified in the literature review, the clustering technique is proposed to develop corruption risk profiles and to use the "item response theory to extrapolate from observed characteristics to latent corruption risks" [6]. Besides the classification of observations, certain studies have aimed at identifying anomalies [16]. Process mining has proved that procedures with more than one bid do last longer and that some single bid tenders lasted an extremely short period of time [31]. This segment raises the question of further analysis of the detected anomalies by the expert, all with the aim distinguishing whether the results are fraudulent. In addition to data-driven by companies, one of the future ideas is that, instead of analyzing Economic Operators, contracts need to be analyzed, which requires a lot of work in some countries because such contract databases are not related, or they don't exist. All processes at the end are governed by humans, and one of the studies showed that bureaucrats that are less reliant on political connections reduces the risks of corruption [39]. The final state is that the sources of data are rather scarce, which greatly affects the outcome of the classification itself [1,18], while on the other hand there is an opinion that there is a possibility of expanding models focusing precisely on economic operators, but with a risk management process approach to creating government services [21,22].

The digitization of the public procurement process certainly offers fewer opportunities for manipulating the process itself, but it is still necessary to increase the efficiency of the fight against corruption in public procurement by enforcing the law and making better use of government resources [6,41].

V. Conclusion

Detection of public procurement corruption in recent years has become one of the major issues around the world. The number of services and amount of money that goes through public procurement is quite large, and for this reason, it is necessary to detect and stop any form of corrupt behavior. Various authors, through various techniques and methods, have been trying to create models that will find any suspicious public procurement process, contract, or economic operator, or classify observations as corrupt or suspicious (Table IV). Of course, this is only one part of the goals that were identified in this paper.

Furthermore, the problem is that there is very little information on PPP that is defined as corrupt, which is a challenge in the techniques that learn from historical data. For this reason, researchers have introduced concepts such as suspicious, bad, or risky PPPs and thus marked the transactions. Data mining and machine learning methods, such as logistic and linear regression, neural networks, process, and text mining, etc. are used in this segment over a large amount of data collected from different data sets, such as contract registers, blacklist economic operators, business registers and so on [1,18,21]. In addition to classification techniques, with the aim of detecting connections between economic operators and contracting authorities, but also for finding daughter companies that participated in collusion of PPP, associations rules and graph databases algorithms were used.

The used metrics are related to the methods, so the most-used metrics in the area of classification or prediction are accuracy, recall, and precision, but unfortunately, this is not the case in all thematic studies, so it is difficult to make a true comparison only with the accuracy metric [4,16,24,26,34]. Moreover, the results obtained vary and depend on the quality

of the data. Much effort has been invested in detecting quality corruption indicators or attributes that have a particular connection to any form of suspect, bad, or risky transactions. Detection takes place at all stages of the public procurement process, from the pre-tender phase to the awarding and post-award phases, but the focus is on using the model as early as possible to prevent a loss of public money or the making of an early warning system. Unfortunately, it has been noted in works that such advanced systems have been integrated into only a small number of state agencies, such as the CGU (Brazilian Office of the Comptroller General) [13,16,21,25,26]. For this reason, the authors point out numerous open issues and suggest combining different methods to improve public procurement processes.

The most effective actions are identification, ranking, and addressing all risks among the ecosystem [5]. Policymakers should perform robust risk assessments, gathering internal input from participants across the ecosystem and across geographies to identify risks and assess mitigating factors. These assessments should also incorporate external factors. There is a wealth of information available in the public domain, and ignoring it results in a big miss. Risks should be assessed at regular intervals (not through a "one and done" approach). Technology should be backed up with appropriate governance, expertise, and monitoring. One single tool won't address all fraud, and technology alone won't keep the process in place. Technology is often only as good as the expert resources, data management and visibility, robust controls, and regular monitoring dedicated to it. Finally, one of the most important actions is being able to react to fraud once identified. This is critical and is a foundational element of an effective fraud policy. The ability to quickly engage the right combination of people, processes, and technology can limit the potential damage. Disruptive fraud often disguises a strategic inflection point, triggering the opportunity for broader social transformation.

REFERENCES

[1] J. Gallego, G. Rivero, J. D. Martínez, "Preventing rather than Punishing: An Early Warning Model of Malfeasance in Public Procurement," Documentos de trabajo 016724, Universidad del Rosario, 2018.

[2] European Court of Auditors, "Fighting fraud in EU spending: action needed," [Online]. Available: http://publications.europa.eu/webpub/eca/special-reports/fraud-1-2019/en/.

[3] K.S. Azmi, A.A. Rahman, "E-Procurement: A Tool to Mitigate Public Procurement Fraud in Malaysia?," In European Conference on Digital Government, Academic Conferences International Limited, Jun 2015, (p. 361).

[4] I. Schuster, S. Merjan, "Assessment Report of corruption risks in public procurement in the Republic of Moldova," Report under project Strengthening the corruption prevention and analysis functions of the National Anti-corruption Center, 2016.

[5] W. Wensink, J. Maarten de Vet, "Identifying and Reducing Corruption in Public Procurement in the EU," 2006, [Online]. Available: https://ec.europa.eu/anti-fraud/sites/antifraud/files/docs/body/identifying_reducing_corruption_in_public_procurement_en.pdf.

[6] J. Ferwerda, I. Deleanu, B. Unger, "Corruption in public procurement: finding the right indicators," European Journal on Criminal Policy and Research, Jun 2017, 1;23(2):245-67.

[7] European Commision, "Single Market Scoreboard, Public Procurement", [Online]. Available: http://ec.europa.eu/internal_market/scoreboard/performance_per_policy_area/public_procurement/index_en.htm.

[8] World Bank, "Enhancing Government Effectiveness and Transparency: The Fight Against Corruption. World Bank," Kuala Lumpur, World Bank. https://openknowledge.worldbank.org/handle/10986/34533 License: CC BY 3.0 IGO., 2020.

[9] Association of Certified Fraud Examiners, "Report to the Nations: 2020 Global Study on Occupational Fraud and Abuse," [Online]. Available: https://www.acfe.com/report-to-the-nations/2020/.

[10] PricewaterhouseCoopers, "Fighting fraud: A never-ending battle PwC's Global Economic Crime and Fraud Survey," [Online]. Available: https://www.pwc.com/gx/en/forensics/gecs-2020/pdf/global-economic-crime-and-fraud-survey-2020.pdf.

[11] E. Dávid-Barrett, "Methods for Corruption in Public Procurement: A Review of Theory and Methodologies," IPA 2011 - Reinforcing Support of CSOs' in Enhancing Transparency and Good Governance in Croatian Public Administration, 2015.

[12] B. K. Tas, "Collusion Detection in Public Procurement with Limited Information," Available at SSRN 2929222. Mar 2017.

[13] R. A. Baldomir, G. C. Van Erven, C. G. Ralha, "Brazilian Government Procurements: an Approach to Find Fraud Traces in Companies Relationships," In Anais do XV Encontro Nacional de Inteligência Artificial e Computacional, Oct 2018, (pp. 752-762). SBC.

[14] R. J. Bolton, D. J. Hand, "Statistical fraud detection: A review," Statistical science, Aug 2002, 1:235-49.

[15] R. S. Michalski, J. G. Carbonell, T. M. Mitchell, "Machine Learning An Artificial Intelligence Approach," Tioga, Paolo Alto,CA, 1983.

[16] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, G. N. Ramos, "Identifying IT purchases anomalies in the Brazilian government procurement system using deep learning," In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Dec 2016, (pp. 722-727).

[17] E. L. Mencıa, S. Holthausen, A. Schulz, F. Janssen, "Using data mining on linked open data for analyzing e-procurement information," In Proceedings of the first DMoLD: Data Mining on Linked Data Workshop at ECML/PKDD, 2013.

[18] Y. Wang, "Detecting Fraud in Public Procurement," Abstract of the Dissertation, University Libraries on behalf of The Graduate School at Stony Brook University, 2016.

[19] R. N. Carvalho, L. J. Sales, H. A. Da Rocha, G. L. Mendes, "Using Bayesian networks to identify and prevent split purchases in Brazil," In Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications, Workshop-Volume 1218 Jul 2014, (pp. 70-78).

[20] L. J. Sales, R. N. Carvalho, "Measuring the Risk of Public Contracts Using Bayesian Classifiers," InBMA@ UAI, Jun 2016, (pp. 7-13).

[21] L. Sales, "Risk prevention of public procurement in the brazilian government using credit scoring," OBEGEF-Observatório de Economia e Gestão de Fraude & OBEGEF Working Papers on Fraud and Corruption; 2013 Jan.

[22] T. Sun, L. J. Sales, "Predicting public procurement irregularity: An application of neural networks," Journal of Emerging Technologies in Accounting. 2018;15(1):141-54.

[23] M. Fazekas, G. Kocsis, "Uncovering high-level corruption: cross-national objective corruption risk indicators using public procurement data," British Journal of Political Science. 2020 Jan;50(1):155-64.

[24] M. Fazekas, I. J. Tóth, L. P. King, "An objective corruption risk index using public procurement data," European Journal on Criminal Policy and Research. 2016 Sep 1;22(3):369-97.

[25] C. G. Ralha, C. V. Silva, "A multi-agent data mining system for cartel detection in Brazilian government procurement," Expert Systems with Applications. 2012 Oct 15;39(14):11642-56.

[26] G. C. Van Erven, R. N. Carvalho, M. T. de Holanda, C. Ralha, "Graph database: A case study for detecting fraud in acquisition of brazilian government," In 2017 12th Iberian Conference on Information Systems and Technologies (CISTI) 2017 Jun 21 (pp. 1-6).

[27] H. A. Arief, G. A. P. Saptawati and Y. D. W. Asnar, "Fraud detection based-on data mining on Indonesian E-Procurement System (SPSE)," 2016 International Conference on Data and Software Engineering (ICoDSE), Denpasar, 2016, pp. 1-6.

[28] N. Modrušan, K. Rabuzin, L. Mršić, "Improving Public Sector Efficiency using Advanced Text Mining in the Procurement Process," In Proceedings of the 9th International Conference on Data Science, Technology and Applications - Volume 1: DATA, 2020, pages 200-206.

[29] K. Rabuzin, N. Modrusan, "Prediction of Public Procurement Corruption Indices using Machine Learning Methods," In KMIS 2019 (pp. 333-340).

[30] F. Decarolis, C. Giorgiantonio, "Corruption red flags in public procurement: new evidence from Italian calls for tenders," Questioni di Economia e Finanza, Occasional Papers. 2020 Feb 1(544).

[31] K. Rabuzin, N. Modrusan, S. Krizanic, R. Kelemen, "Process Mining in Public Procurement in Croatia," 8th International Scientific Conference on Industrial Systems Industrial Innovation in Digital Age, Novi Sad 2020, in press.

[32] R. B. Velasco, I. Carpanese, R. Interian, O. C. Paulo Neto, O. C. Ribeiro, "A decision support system for fraud detection in public procurement," International Transactions in Operational Research. Jan 2021;28(1):27-47.

[33] DG GROW, "Study on up-take of emerging technologies in public procurement," [Online]. Available: https://joinup.ec.europa.eu/sites/default/files/news/2020-06/D.01.06_Final_report_v3.00.pdf.

[34] M. Sokolova, N. Japkowicz, S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," In Australasian joint conference on artificial intelligence, Springer, Berlin, Heidelberg, Dec 2006 (pp. 1015-1021).

[35] P. Cichosz, "Data mining algorithms: explained using R," John Wiley & Sons Incorporated; Jan 2015.

[36] "Performance Measurement in Public Procurement," [Online]. Available: https://www.publicspendforum.net/blogs/psfeditorial/2019/04/02/performance-measurement-public-procurement/.

[37] OLAF, "The OLAF report 2019," The Publications Office of the European Union, Luxembourg, [Online]. Available: https://ec.europa.eu/anti-fraud/sites/antifraud/files/olaf_report_2019_en.pdf.

[38] Corruption perceptions index, Transparency International, [Online]. Available: https://www.transparency.org/news/feature/coruption_perceptions_index_2017.

[39] N. Charron, C. Dahlström, M. Fazekas, V. Lapuente, "Careers, Connections, and Corruption Risks: Investigating the impact of bureaucratic meritocracy on public procurement processes," The Journal of Politics. Jan 2017;79(1):89-104.

[40] A. Dhurandhar, T. Ravi, B. Graves, G. Maniachari, M. Ettl, "Robust system for identifying procurement fraud," In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Jan 2015, (pp. 3896-3903).

[41] A. Afolabi, E. Ibem, E. Aduwo, P. Tunji-Olayeni, "Digitizing the grey areas in the Nigerian public procurement system using e-Procurement technologies," International Journal of Construction Management. 2020 Jun 4:1-0.

[42] Y. T. Berru, V. F. L. Batista, P. Torres-Carrión, M. G. Jimenez, "Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review," In International Conference on Applied Technologies, Springer, Cham, Dec 2019, (pp. 254-268).

[43] D. Carneiro, P. Veloso, A. Ventura, G. Palumbo, J. Costa, "Network Analysis for Fraud Detection in Portuguese Public Procurement," In International Conference on Intelligent Data Engineering and Automated Learning, Springer, Cham, Nov 2020 (pp. 390-401).

Mobile-based Decision Support System for Poultry Farmers: A Case of Tanzania

Martha Shapa[1], Lena Trojer[2], Dina Machuve[3]
School of Computational and Communication Science and Engineering
Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania[1, 3]
Department of Technology and Aesthetics Blekinge Institute of Technology, Karlshamn, Sweden[2]

***Abstract*—Poultry farms in Tanzania are characterized by inadequate management practices which are mainly caused by the lack of adequate systems to guide the small-scale poultry farmers in decision making. It is well-established that information is a key factor in making effective decisions in numerous sectors including poultry farming. Furthermore, various researchers have identified the use of mobile decision support tools to be an effective way of aiding farmers in making informed decisions. In this paper, we present a mobile-based decision support system that will aid rural and small-scale poultry farmers in Tanzania to obtain reliable information that is crucial for making proper decisions in their farming activities. In this context, a mobile-based decision support system was achieved through a mobile application integrated with a chatbot assistant to provide a solution to various poultry farming-related problems and simplify their decision-making process. We used a data-driven approach towards developing an informational chatbot assistant for Android smartphones that is capable of interacting with small-scale poultry farmers through natural conversations by utilizing the RASA framework.**

Keywords*—Decision support system; chatbot; mobile application; poultry farming; data-driven approach

I. INTRODUCTION

Poultry farming is one of the prominent agricultural sectors dealing with the keeping of various domestic birds to produce eggs and meat for consumption and trade [1]. Similarly, it is one of the important agricultural areas for generating income for farmers in Tanzania. Studies indicate that there are approximately 36.2 million chickens in Tanzania, of which approximately 95% are local chickens reared by rural households [2]. Over the past decade, the growth of poultry production has accelerated, due to the rapid urbanization and increase in demand for poultry products, which include chicken meat and eggs [3]. Poultry farming has a direct impact on the farmers and has gained a notable attraction among entrepreneurs and women. Generally, it is the source of the poultry farmer's family income and protein. In this sector, women constitute the majority of poultry farmers as they constitute over 80% of the farming population in Tanzania [4].

The majority of small-scale farmers in Tanzania rely on poultry farming as their major source of income [5]. However, poultry production is hindered by several challenges including unreliable markets, poultry diseases, scarce inputs, and shortage of timely extension information due to scarcity of extension officers as well as distant locations for consultation [1]. It has been observed that rural and small-scale poultry farmers rely mainly on unreliable sources of information for poultry management such as word of mouth from family members, neighbors, and friends with previous poultry keeping experience due to the lack of adequate systems to guide them in decision making [6]. Moreover, information is very important in the development of poultry farming and agriculture at large, therefore the information obtained from unreliable sources may lead to underdevelopment of this sector, especially to rural, peri-urban, and small-scale poultry farmers [1].

Technology advancement plays a great role in the agricultural sector development, including poultry farming. It has been argued that the use of mobile decision support tools is an effective way of aiding farmers to make informed decisions [6]. Due to an increase in the use of mobile communication technologies in different sectors in the country [7], this study will contribute and improve proper information attainment for poultry farming, and aiding farmers to make informed decisions.

Along with the technological advancement and the increase in the information-seeking behavior of the small-scale poultry farmers in Tanzania [6], this study will play a great role in aiding the small-scale poultry farmers attain crucial poultry farming information in time and make informed decisions. A mobile-based decision support system is achieved through a mobile application with a chatbot assistant that provides a solution to various poultry farming-related problems and simplifies their decision-making process. Furthermore, the conversational assistant, chatbot, is a modern human-computer interaction technology that was introduced in the 1960s when the earliest chatbot ever was developed [8], [9], and gained popularity in 2016. It has been argued that chatbot is one of the most advanced and effective ways to provide information and facilitate the decision-making process in various sectors [10]. According to [8], [11], chatbots can be used to aid farmers by providing information and solutions through responding to poultry farming-related problems and facilitate their decision-making process in poultry farming.

This paper introduces a decision support mobile application, for providing poultry farming-related information to small-scale poultry farmers in Tanzania. For this purpose, the mobile application was developed using Android Studio, and integrated with a chatbot.

The remainder of this paper is organized as follows: Section II presents the related works carried out in solving

various farm-related problems. Section III demonstrates the methodology used in the requirements gathering, the approach and tools used in the development of the mobile-based Decision Support system. Section IV discusses the results of the proposed system, and followed by conclusion in Section V.

II. Related Works

Various researchers have revealed that the use of mobile-based conversational assistants is an effective way to aid farmers with farm-related information and problem-solving. In the early days after the first chatbot was developed, chatbots had limited effectiveness and maintained a simple conversational flow. As the research progresses, recent chatbots are capable of understanding the context of the user and the flow of the conversations and provide a suitable response. Several studies have been conducted to develop chatbots that will assist farmers in solving their farm-related problems.

Jain et al. [11] designed a chatbot called FarmChat that aims to meet the information needs of farmers in rural India. The system offers information to the farmers by answering their farming-related queries. It was developed using the IBM Watson Assistant and consists of two interface modalities: Audio-only, and Audio+Text [11]. The study was conducted with 34 potato farmers in rural India and indicated that the chatbot offered satisfying information that supported them [11]. Thus, the authors suggested that conversational assistance delivered through smartphones could be an effective way to improve the information accessible to people with limited literacy in rural areas.

Arora et al. [8] developed an interactive chatbot named Agribot, that assists farmers in problem-solving, crop disease detection, and weather prediction. They developed the chatbot using sequence-to-sequence learning, an approach that allows the model to learn the mapping between questions and their suitable response [8]. The authors suggested that the chatbot could be more generalized in terms of conversations if the model is trained in a massive amount of data-points [8].

Fue et al [12] developed an agro-advisory web and mobile-based system called 'Ushaurikilimo' that allows farmers to request advisory services from an agriculture extension officer using either the web or mobile phone [12]. 'Ushaurikilimo' operates in Tanzania. It is a two-way communication platform between farmers and experts [12]. It functions by allowing farmers to ask questions through SMS and get a response from the agricultural expert [12]. The platform allows farmers to ask for advice on agricultural-related issues like; farm management, livestock keeping, marketing information, and aquaculture [12], whereby it depends on the presence of the experts in order for the farmers' problems to be solved. The proposed system aims at solving this gap.

III. Methodology

The methodological approach we used in the development of a mobile-based Decision Support system for small-scale poultry farmers in Tanzania in this study is the utilization of Android Studio in the mobile application development and a Rasa framework in the development of a chatbot that will aid small scale poultry farmers in Tanzania by giving answers to their poultry-related problems and help them make an informed decision in their poultry management practices.

As illustrated in Fig. 1, when the small-scale poultry farmer types his or her poultry-related query in the mobile application's chat window and sends it, the text is fed into Rasa NLU through the Application Programming Interface (API). An API software intermediary allows the farmer's mobile application and our assistant to communicate by delivering the requests from the poultry farmer to the chatbot assistant and delivering the response back to the poultry farmer [13]. After the Rasa NLU receives the text message from the poultry farmer in form of a natural human language, performs intent classification, entity extraction, and converts it into the form of structured data that our chatbot assistant could understand what the farmer is saying.

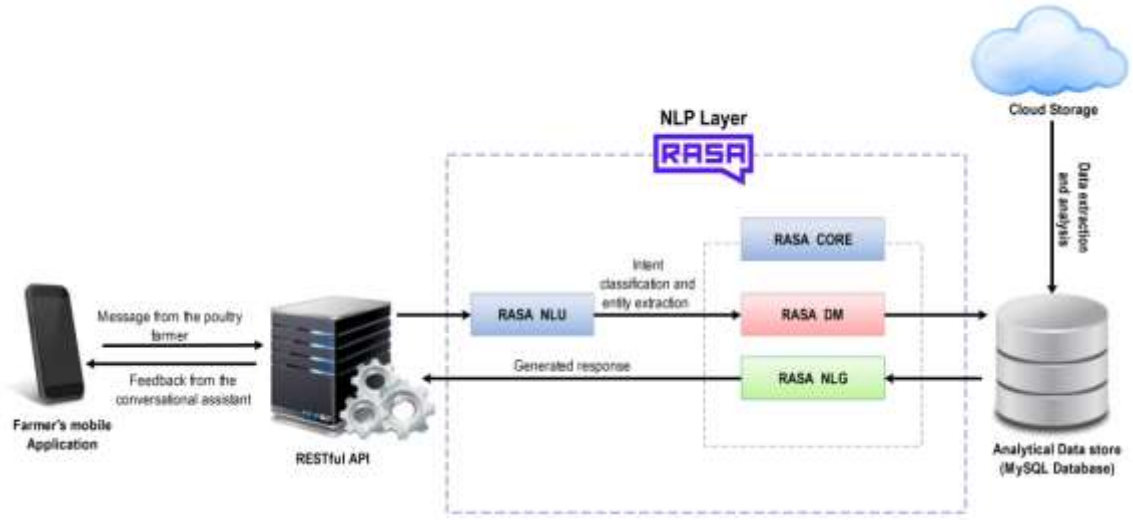


Fig. 1. A Proposed Conceptual Framework

Furthermore, based on the intents and entities, the Rasa Core takes structured inputs from the Rasa NLU through the dialogue management and predicts the next desirable action using a probabilistic model called Long short-term memory (LSTM) neural network [14], through the LSTM based supervised learning (SL) and Reinforced learning (RL) rather than the if/else statement [14]. Reinforced Learning (RL) is used to improve the next best action. Therefore, Rasa Core performs Dialogue Management by keeping track of the conversation and decide the next set of actions to be performed by the chatbot [15]. An action can be a simple utterance that means sending a feedback message to the poultry farmer, or an arbitrary function to execute. After the action has been executed, it passes a tracker instance to use any relative information collected in the dialogue history and previous actions [15].

Additionally, Rasa Core has high Natural Language Generation (NLG) capabilities [16], [17]. Therefore, upon retrieval, the Rasa Core uses its NLG capabilities to prepare a natural language human-like response to the poultry farmer based on the intent and context information returned from the Rasa NLU [17]. The dialogue manager generates raw responses that will be passed to the Natural Language Generator component that refines the text response and construct the understandable text responses in natural human language in machine representation. NLG process converts structured data into text, therefore it generates an appropriate response that a human can understand [16]. Finally, the generated feedback is sent back to the farmer's mobile application User interface via the API.

A. Data Collection and Requirements Gathering

An exhaustive literature review was conducted to identify and assess the information management requirements of small-scale poultry farmers. The identified information management requirements as summarized in Table I. Includes literature, frequency, and percentage of frequently asked questions by poultry farmers in attempts to attain proper information for their proper poultry management practices. The identified information management requirements of poultry farmers according to Table I, includes:

1) Chicken health: From an exhaustive literature review, various studies have identified that the majority of the small-scale poultry farmers in Tanzania are facing challenges in maintaining the good health of their poultry [6]. The studies show that small-scale poultry farmers in Tanzania face many health-related problems in their poultry farms and hence tend to seek various information concerning poultry health from various sources of information, mostly unreliable sources. The health information they mostly seek includes; disease control, diagnosis of the chicken diseases, the transmission of diseases, and vaccination of chickens against different diseases [1], [6].

2) Chicken feeds: One of the most important information needs that small-scale poultry farmers in Tanzania seek includes chicken feeds information [1]. The small-scale poultry farmers frequently ask questions about the availability of the chicken feeds, Types of feeds for their chicken types, amount of feeds per chicken, and feed formulation and preparation for the chicken feed, to maintain the proper production [6].

3) Chicken breeds: According to [6], poultry farmers are interested in knowing the different types of chicken breeds for various purposes before starting up a poultry farm. The factors that the majority of small-scale poultry farmers consider include; diseases resistant breed types, breed types that are best for commercial purposes, and breed types for egg production [6]. The poultry farmers also seek general information on chicken breeds to help them in proper keeping of a particular chicken breed, and rearing techniques of different types of chicken breeds.

4) Egg production: Among the frequently asked questions by many small-scale poultry farmers in Tanzania includes questions about egg production, particularly for commercial purposes. [6], [18] highlights that small-scale poultry farmers seek information about egg production, which includes improving the quality and production of eggs, proper storage of eggs, and need to know the best method for incubation of eggs [6].

5) Housing: Various studies have highlighted that majority of the startup small-scale poultry farmers tend to seek information on the housing of the chicken shelter before they start practicing poultry farming [6]. The farmers are more interested in attaining the information about the startup capital for poultry farming, characteristics of the chicken house and how to build it regarding the geographical position of the farm, and size of the chicken house for a certain number of chickens [1].

The identified information management requirements of small-scale poultry farmers through literature review were the requirements gathered and used in the development of the mobile application for poultry farming data collection, and training of our chatbot assistant to offer reliable poultry farming information that will enable the small-scale poultry farmers to make informed decisions. The information management requirements gathered helped in the identification and attainment of the 200 sample questions and answers for poultry farming-related problems, these sample questions and answers were used in training the NLU model of our chatbot assistant.

B. Approach for Mobile Application Development

This study aims at the development of a mobile-based decision support system, which involves the development of a mobile application. The approach used in the development of the mobile application is the Android studio with JAVA programming language [20]. This approach was chosen due to the reasons that the Android mobile operating systems are widely used by the targeted users [21]. Android Studio was selected because it is suitable for the development of Android-based applications for smartphones with Android operating systems [20]. The minimal version of the Android operating system supported by the developed mobile application is Jelly Bean 4.3, this was selected because it is more inviting for the average user.

TABLE I. INFORMATION MANAGEMENT REQUIREMENTS FOR POULTRY FARMERS

Number	**Literature Author**	**Sample Space**	**Information need**	**Frequency**	**Percentage**
1.	Grace Msoffe et al. [6]	187	Disease control	187	100
			Breeds and breeding	78	41.7
			Housing and shelter	56	29.9
			Feeding and nutrition	47	25.1
2.	Temba B et al. [1]	160	Poultry diseases	112	70
			Poultry nutrition	92	57.5
			Housing	92	57.5
3.	Benjamin Folitse et al. [18]	150	Disease management	145	96.7
			Eggs production	114	76
			Feeding and nutrition	112	74.7
			Shelter	108	72
4.	Jotshana Khobragade et al. [19]	60	Health and disease control	56	93.34
			Feeding management	55	91.67
			Housing management	53	88.34

C. Natural Language Processing Layer

The methodological approach we used in this study is the utilization of a Rasa framework in the development of a chatbot that will aid small scale poultry farmers in Tanzania by giving answers to their poultry-related problems and help them make an informed decision in their poultry management practices. The Rasa framework is an open-source machine learning framework for building contextual conversational assistants called chatbots, these assistants consist of two components which are Rasa NLU and Rasa Core [22].

1) Rasa NLU: This is the Rasa's desirable library for Natural Language Understanding that performs intent classification and entity extraction [17]. It takes user inputs in a simple unstructured human language and extracts structured semantic information in the form of intents and entities [17], [23]. Intents are labels that are attached to each user's input based on the overall goal of the user's message, and entities are pieces of information that our conversational assistant may need in a certain context. Furthermore, Rasa NLU is treated as the ear of the chatbot, because it teaches the chatbot to understands the inputs of the user [22].

2) Rasa Core: This is a framework for machine learning-based contextual decision making so-called the brain of our assistant because it predicts how our assistant will respond based on a specific state of the conversation as well as the context [17], [22]. It learns by observing the pattern from example conversational data between the user and the assistant also called stories.

Rasa Core is responsible for Dialogue management (Rasa DM). In Rasa, Dialogue management learns the patterns of the conversations from the example conversational data using Machine learning and predict how our assistant should respond in a specific situation based on the history of the conversation and the context [15], [17]. In Dialogue management the training Data for our conversational assistant is called stories, these are example conversations between the small-scale poultry farmer and our assistant, written in a specific format. This format includes expressing the user inputs as relative intents and entities, the same way as they were expressed in the NLU training Data, while the responses of our assistant were expressed as action names [15].

Furthermore, Rasa Core has high NLG capabilities that enable the chatbot to intelligently know the exact and clear response that is to be generated for a corresponding user message [17].

D. Training of the Rasa Conversational System Model

In the training of our conversational model, both Rasa NLU and Rasa Core use human-readable training data formats. Rasa NLU requires a list of utterances that are annotated with intents and entities for training our chatbot assistant [15]. We used both JSON structure and markdown format in the training of our chatbot assistant [15]. Using the Rasa NLU pipeline, we prepared a training data set to classify the intents and extract the entities. The training data includes several intents: greet, goodbye, chicken_feeds, chicken_breeds, eggs_production, chicken_diseases, and chicken_shelter. We use about 200 sample questions and answers with marked entities to train the Rasa NLU [15], [16].

In addition to the supervised learning, Rasa Core supports a machine teaching approach whereby, the actions made by the system can be corrected by the developers, we used this approach in generating training data and inspecting the space of credible conversations efficiently [15], [24]. The Training data used in training the Rasa Core are known as stories, these are the sample conversations between the user and our chatbot assistance [24]. Furthermore, Rasa core's Machine Learning libraries give it the capability of learning from the previous conversations between the user particularly the poultry farmer, and our chatbot assistant [17].

1) Markdown training data format example

```
## intent:chicken_diseases
- How do you treat a chicken’s skin wounds?"
- Can I keep sick poultry together with normal ones?
- What is causing your hen’s swollen foot and her limping?

## intent:egg_production
- How often do [layers](flock_name) lay eggs?
- When will my hens start laying?
- How long do chickens lay eggs?
- Why do some eggs have soft shells or no shells?

## intent:greet
- hi
- hey
- hello
```

2) JSON training data format example

```
"rasa_nlu_data": {
    "common_examples": [
        {
            "text": "Hello",
            "intent": "greeting",
            "entities": []
        },
        {
            "text": "How do you treat a chicken’s skin wounds?",
            "intent": "chicken_diseases",
            "entities": []
        },
        {
            "text": " Can I keep sick poultry together with normal
ones?",
            "intent": " chicken_diseases",
            "entities": []
        },
        {
            "text": " When will my hens start laying?",
            "intent": "egg_production"
        }
    ],
    "regex_features": [],
    "entity_synonyms": []
    }
}
```

IV. Results and Discussion

Based on the study described, a mobile-based Decision Support system for poultry farmers was implemented. The developed Android-based mobile application for poultry-related data storage and poultry farming-related information provision was integrated with our chatbot assistant, for the aim of aiding small-scale poultry farmers with reliable information for productive management practices.

The developed mobile application as illustrated in Fig. 2, consists of five modules; Consultation module, Information portal, events, new records, and my records. The information portal contains various general information that the small-scale poultry farmer will require in poultry farming.

The new records module allows the small-scale poultry farmer to keep records of his or her day to day flock management activities, feed management, which includes the amount and type of feeds the farmer offers to the chicken. Medication records, the farmer will be able to keep records of all medications provided to the chickens, which includes the vaccination record, vitamins provision, and general medications offered to the farmer’s flock, lastly are the records of the finances, the farmer can keep records of the sales and expenditures of the poultry farm. The farmer can view the farming records and keep track of the development of the poultry farm on a timely basis.

The recorded poultry farming data are stored in the cloud storage, together with the stored previous conversations are used by our chatbot assistant in responding to various poultry-related questions, and advice the farmer regarding the particular poultry farmer’s farm and flock condition. This Data-driven Approach used in the development of our mobile-based decision support system makes our system intelligent enough to help the poultry farmers in their decision making for productive poultry management practices [25].

The developed mobile application consists of a consultation module. The consultation module is the chatbot assistant that responds to the poultry farmers' questions concerning chicken health, chicken feeds, chicken eggs production, and the chicken breeds. The chatbot assistant was trained to offer consultation to the small-scale poultry farmers regarding the most common poultry-related problems that mostly face them, and help the poultry farmer make proper decisions in practicing poultry farming. The consultation chat between the user and our chatbot assistant is illustrated in Fig. 3.

Fig. 2. Poultry Farmer's Mobile Application.

Fig. 3. Conversations on the Consultation Module.

V. Conclusion

This paper has presented an exploratory study using a mobile conversational agent-based interaction to facilitate intelligent decision support to the small-scale poultry farmers during a consultation with our assistant. We implemented the proposed mobile-based decision support system for poultry farmers via an interactive chatbot assistant using a Rasa framework, and an Android-based mobile application using Android studio.

Future work focuses on the validation of the mobile application developed to ensure the extent to which this study contributes in aiding farmers with the crucial poultry farming information that they frequently seek for making informed decisions. This process will involve user acceptance testing, to testing, by letting the small-scale poultry farmers use the mobile application by storing their poultry-related data and consult our chatbot assistant via the mobile application.

ACKNOWLEDGMENT

The authors would like to acknowledge the support in this study from the Organization for Women in Science for the Developing World (OWSD) Early Career Fellowship Programme (Agreement number 506575) and the African Development Bank (AfDB).

REFERENCES

[1] B. A. Temba, F. K. Kajuna, G. S. Pango, and R. Benard, "Accessibility and use of information and communication tools among farmers for improving chicken production in Morogoro municipality, Tanzania," Livest. Res. Rural Dev., vol. 28, no. 1, 2016.

[2] A. Saleque, A. Jabeen, and S. M. Real, "Small Scale Poultry Rearing in Tanzania – Subsistence to Surplus Production for Increase Income and Improve Food and Nutrition Security.," pp. 1–7, 2016.

[3] A. Vernooij, M. N. Masaki, and D. Meijer-Willems, "Regionalisation in poultry development in Eastern Africa," 2018, [Online]. Available: www.wageningenUR.nl/livestockresearch.

[4] E. L. Isaya, R. Agunga, and C. A. Sanga, "Sources of agricultural information for women farmers in Tanzania," Inf. Dev., vol. 34, no. 1, pp. 77–89, 2018, doi: 10.1177/0266666916675016.

[5] A. Ali et al., "Backyard poultry farming empowering women for doubling farmers ' income," vol. 8, no. 2, pp. 1138–1143, 2020.

[6] G. Msoffe, A. Chengula, M. J. Kipanyula, M. R. S. Mlozi, and C. A. Sanga, "Poultry Farmers' Information needs and Extension advices in Kilosa, Tanzania: Evidence from Mobile-based Extension, Advisory and Learning System (MEALS)," Libr. Philos. Pract., vol. 2018, no. February, 2018.

[7] BMI, "Tanzania Telecommunications Report," p. 62, 2016.

[8] B. Arora, D. S. Chaudhary, M. Satsangi, M. Yadav, L. Singh, and P. S. Sudhish, "Agribot: A Natural Language Generative Neural Networks Engine for Agricultural Applications," 2020 Int. Conf. Contemp. Comput. Appl. IC3A 2020, pp. 28–33, 2020, doi: 10.1109/IC3A48958.2020.233263.

[9] M. H. Wen, "A conversational user interface for supporting individual and group decision-making in stock investment activities," Proc. 4th IEEE Int. Conf. Appl. Syst. Innov. 2018, ICASI 2018, pp. 216–219, 2018, doi: 10.1109/ICASI.2018.8394571.

[10] O. Chukhno and N. Chukhno, "A chatbot as an environment for carrying out the group decision making process," vol. i, pp. 15–25, 2019.

[11] M. Jain, P. Kumar, I. Bhansali, Q. V. Liao, K. Truong, and S. Patel, "FarmChat: A Conversational Agent to Answer Farmer Queries," Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., vol. 2, no. 4, pp. 1–22, 2018, doi: 10.1145/3287048.

[12] K. Fue, A. Geoffrey, M. Mlozi, S. Tumbo, R. Haug, and C. Sanga, "Analyzing usage of crowdsourcing platform Ushaurikilimo' by pastoral and agro-pastoral communities in Tanzania," 2016.

[13] R. Gunawan, I. Taufik, E. Mulyana, O. T. Kurahman, M. A. Ramdhani, and Mahmud, "Chatbot Application on Internet of Things (IoT) to Support Smart Urban Agriculture," Proceeding 2019 5th Int. Conf. Wirel. Telemat. ICWT 2019, pp. 3–8, 2019, doi: 10.1109/ICWT47785.2019.8978223.

[14] Y. Windiatmoko, A. F. Hidayatullah, and R. Rahmadi, "Developing FB chatbot based on deep learning using RASA framework for university enquiries," arXiv, 2020.

[15] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open source language understanding and dialogue management," arXiv, pp. 1–9, 2017.

[16] L. Gunasekara and K. Vidanage, "UniOntBot: Semantic Natural Language Generation based API approach for Chatbot Communication," 2019 Natl. Inf. Technol. Conf. NITC 2019, pp. 8–10, 2019, doi: 10.1109/NITC48475.2019.9114440.

[17] M. Bagchi, "Conceptualising a library chatbot using open source conversational artificial intelligence," DESIDOC J. Libr. Inf. Technol., vol. 40, no. 6, pp. 329–333, 2020, doi: 10.14429/djlit.40.6.15611.

[18] B. Y. Folitse, J. Sam, L. P. Dzandu, and S. K. Osei, "Poultry Farmers' Information Needs and Sources in Selected Rural Communities in the Greater Accra Region, Ghana," Int. Inf. Libr. Rev., vol. 50, no. 1, pp. 1–12, 2018, doi: 10.1080/10572317.2017.1351020.

[19] J. A. Khobragade, V. V. Banthiya, S. P. Landge, A. P. Dhok, M. M. Kadam, and J. D. Nandeshwar, "Information Need of Poultry Farmer from Vidarbha Region of Maharashtra," Int. J. Curr. Microbiol. Appl. Sci., vol. 8, no. 02, pp. 3373–3378, 2019, doi: 10.20546/ijcmas.2019.802.392.

[20] A. Sarkar, A. Goyal, D. Hicks, D. Sarkar, and S. Hazra, "Android Application Development : A Brief Overview of Android Platforms and Evolution of," 2019 Third Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), pp. 73–79, 2019.

[21] S. Saengwong and C. Koksantia, "Development of a mobile app for enhancing the performance of smallholder native chicken management and production," Asia-Pacific J. Sci. Technol., vol. 25, 2020.

[22] Rakesh Kumar Sharma, "An Analytical Study and Review of open source Chatbot framework, Rasa," Int. J. Eng. Res., vol. V9, no. 06, pp. 1011–1014, 2020, doi: 10.17577/ijertv9is060723.

[23] D. Braun, A. H. Mendez, F. Matthes, and M. Langen, "Evaluating natural language understanding services for conversational question answering systems," SIGDIAL 2017 - 18th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue, Proc. Conf., no. August, pp. 174–185, 2017, doi: 10.18653/v1/w17-5522.

[24] K. Deepika, V. Tilekya, J. Mamatha, and T. Subetha, "Jollity Chatbot- A contextual AI Assistant," Proc. 3rd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2020, no. Icssit, pp. 1196–1200, 2020, doi: 10.1109/ICSSIT48917.2020.9214076.

[25] S. Hwang, B. Kim, and K. Lee, A Data-Driven Design Framework for Customer Service Chatbot, vol. 11583 LNCS. Springer International Publishing, 2019.

Using Behaviour-driven Requirements Engineering for Establishing and Managing Agile Product Lines

An Observational Study

Heba Elshandidy[1], Sherif Mazen[2], Ehab Hassanein[3], Eman Nasr[4]
Information Systems Department
Faculty of Computers and AI, Cairo University, Cairo, Egypt[1, 2, 3]
Independent Researcher, Cairo, Egypt[4]

***Abstract*—Requirements engineering in agile product line engineering refers to both common and variability components establishing a software. Although it is conventional for the requirements engineering to take place in a dedicated upfront domain analysis phase, agile-based environments denounce such a proactive behaviour. This paper provides an observational study examining a reactive incremental requirement engineering approach called behaviour-driven requirements engineering. The proposed approach uses behaviour-driven development to establish and maintain agile product lines. The findings of the study are very promising and suggest the following: the approach is easy to understand and quick to learn; the approach supports the constantly changing nature of software development; and using behaviour-driven requirements engineering produces reliable and coherent requirements. In practice, the observational study showed that using the proposed approach saved time for development team and customers, decreased costs, improved the software quality, and shortened the time-to-market.**

Keywords*—Agile product line engineering; behaviour-driven requirements engineering; observational study; requirements engineering

I. Introduction

Agile product line engineering (APLE) has been gaining a momentum throughout the past decade due to its faster delivery, lesser time-to-market, and more involvement for customers in every development cycle. APLE is the resulting approach of merging agile software development (ASD) and software product line engineering (SPLE); that term was formally coined at the first APLE'06 Workshop [1]. The purpose of APLE is to overcome the weaknesses of both paradigms (i.e., ASD and SPLE) while maximizing their benefits. A software product line (SPL) is a family of software products that share a common set of features (i.e., core assets) in addition to the unique features (i.e., variability) associated to each product in the family that satisfy the different needs of the customers [2]. Thus, it is intuitive to deduce that agile product lines (APLs) are SPLs that are either developed in an entirely ASD environments or in traditional environments that adopt some of the ASD practices. ASD, on the other hand, is a group of incremental and iterative software development methodologies that advocate quick clean software delivery and customers' involvement throughout the project lifetime [3]. The work in this paper focuses on behaviour-driven development (BDD) which is an ASD process that encourages the collaboration between the different stakeholders (i.e., customers, quality assurance, developers, etc.) of a software project [4].

According to the studies in [5,6], there are eleven factors that contribute to the success of a software project. While eight of those factors are related to requirements engineering (RE), ten of them are related to ASD. RE is the process of identifying, analysing, documenting, and managing user requirements [7,8]. The overlapping between the RE-related and the ASD-related project's success factors indicates that they share the same goals. Thus, it is most likely that having an agile-based requirements engineering process highly increases the possibility of having a successful software project.

Having realised the advantages of APLE as a development approach and the critical role of RE in a project's success, it is inquisitive to know whether it is feasible to achieve an incremental agile-based RE approach for APLs using BDD in a real-life empirical case study.

The rest of the paper is organised as follows: Section II explains BDD in further details while Section III briefs the reader about related work. Section IV summarises the proposed behaviour-driven requirements engineering (BDRE) approach. Section V presents the conducted observational study. Section VI discusses the results of the study. Finally, Section VII concludes the paper.

II. Background

BDD was created to overcome the shortcomings of test-driven development (TDD). In particular, the starting point of testing, when and what to test, how much to test, understanding why a test fails, the need to have naming conventions for tests, and knowing whether a specification is met or whether the code delivers a business value [4]. BDD combines the general methods and practices of TDD with concepts from domain-driven design and objected-oriented analysis and design [4]. This provides a shared process and a common understanding to all the involved stakeholders (i.e., developers, designers, etc.). Thus, helps them to successfully collaborate on software development with well-defined outputs. As a result, BDD is capable of delivering working and tested software in shorter time-to-market while better managing traceability between the different artefacts of the system [4].

BDD has six main characteristics [4,9]:

- Ubiquitous language: which is a common language that enables customers and development teams to communicate without ambiguity. That language contains all the terms that will be used to define the behaviour of the systems. Although the structure of such languages emerges from the business domain model, BDD has its own pre-defined domain-independent ubiquitous language.
- Iterative decomposition process: since it is often difficult for the development team to find a starting point through which they can collect the customers' requirements, BDD works in an iterative manner to resolve that issue. Although the customers themselves might not have a clear view of the requirements they need, they surely know the business values and the behaviour they expect from the software project. As a consequence, the analysis process in BDD starts with the identification of the expected behaviour of the system, based on the intended business outcomes, which is later decomposed into a set of features. Each feature is then realised by a set of user stories and each user story is further described through a set of scenarios. A scenario is a specific instance of a particular user story that describes an actual context and output for that user story.
- Plain text description with User Story and Scenario templates: features, user stories, and scenarios are represented in plain text predefined templates using the BDD ubiquitous language. For example, to write a story, the following template is used:

  [UserStoryTitle] (One line describing the story)

  As a [Role]

  I want a [Feature]

  So that [Benefit]

  To write a scenario, the following template is used:

  Scenario 1: [Scenario Title]

  Given [context]

  And [Some more contexts]

  When [Event occurs]

  Then [Outcome]

  And/But [Some more outcomes]

  While a user story describes an activity that is done by a user in a given role, the scenario describes how the system should behave when it is in a specific state for a specific feature and an event happens. Both user stories and scenarios are directly mapped to tests.
- Executable acceptance tests (EATs) with mapping rules: acceptance tests (ATs) in BDD is the satisfaction condition(s) that determines whether the behaviour of a particular feature is successfully achieved. BDD inherits the characteristic of executable testing from automated TDD, where ATs are regarded as automated specifications that verify the behaviour/interaction of the object rather than its state. Mapping rules provide a standardised way of mapping from scenarios to test codes, thus, facilitates managing traceability between the different artefacts of the system.
- Readable behaviour oriented specification code: BDD emphasises the importance of including the code in the system's documentation. Thus, the code should be readable and the specifications should be part of the code. The mapping rules help produce readable behaviour oriented code.
- Cross-cutting through the different software development phases: at the planning phase, the business outcomes are mapped to behaviours, where they are then decomposed into a set of features in the analysis phase. Then at the implementation phase, the EATs take place in which testing classes are derived from scenarios.

III. Related Work

The APLE literature tackled various problems for the different RE activities (i.e., requirements elicitation, analysis, modelling, verification and validation, and management). After thoroughly studying the APLE RE literature and to the best of our knowledge, none of the previous efforts in this area proposed a RE solution that was based on BDD.

Additionally, all the attempts [10-27], except for the efforts in [28-31], focused on adopting ASD practices in already existing SPLE environments. These efforts are placed on the other spectrum of our work which is focusing on building and managing APLE in established agile-based environments.

As a further matter, there were no efforts in the literature that offered a reliable RE solution that addressed the five activities of the RE process. Although there was an all-inclusive RE solution attempt [13,14] in the literature, the authors did not validate their work through either a theoretical or a practical case study. Additionally, the authors collected their data from managers only and disregarded the perspective of the other stakeholders. Thus, directly violates the values of ASD where the perspectives of all the involved stakeholders should be taken into consideration throughout the development lifetime. Finally, none of the literature mentioned in this paper conducted a real-life empirical study to validate the respective proposed work.

The aforementioned research gaps were further confirmed by five systematic literature reviews [32-36]. These studies concluded that RE was not addressed properly or sufficiently in APLE regardless of the agility degree of the used development approach. Based on these findings and in addition to the crucial role of RE in the success of software projects, it has become imperative to have a systematic lightweight RE approach to reactively and incrementally develop and manage APLs.

IV. Summary of the BDRE Approach

The BDRE approach depends on BDD to have an incremental evolutionary flexible RE process. The full details about the BDRE approach are available in [37]. In BDRE, it is assumed that business goals, both functional and non-functional, are already identified and available for the development team to start their RE process. Generally, business goals are derived from the business need of finding solutions for a particular business problem.

The BDRE approach consists of five key activities: requirements elicitation, analysis, modelling, validation and verification, and management. Each activity is briefed as follows:

- Requirements elicitation: This is the first step in the BDRE approach where the work starts outside-in. The input to this activity is the set of solution hypotheses for the already identified business problem. The development team uses prototyping to determine the relevancy of the proposed solutions set to the underlined business goal. After agreeing on the final set of solution, the development team determines the scope of the system accordingly. After that, the development team and the customer's representative decide the initial set of features, reflecting the needed behaviour of the system-under-development (SUD), to be developed in the next iteration. This concludes the elicitation activity with that initial set of features as an output.
- Requirements analysis: This is the second activity in the BDRE approach where the initial user requirements are further examined. The initial set of features from the previous activity in addition to the already existing features, of other products in the same SPL, are fed as an input for the analysis activity. The personnel representing the roles of business analyst, developer, and quality assurance conduct specifications workshops (aka. the three Amigo's meetings) to further analyse and negotiate that given inputs. Firstly, they examine the relevancy and the clarity of the given features in comparison to the business goals. Then, they come to a consensus on which features to consider as core assets and which ones to consider as variabilities. In case they detect an abnormality in the given requirements, they may go back to the requirements analysis activity for further inspection. Otherwise, they conclude this activity by producing an initial set of user stories for each core asset/variability feature.
- Requirements modelling: This is the third step in the BDRE approach with the initial set of user stories, produced at the analysis activity, as an input. The main goal of this activity is to illustrate each user story by an example. This is achieved through developing a series of real scenarios with actual values for each user story. After meetings and negotiations, the development team finalises the initial set of scenarios (i.e., the output of this activity) for each user story of each feature. If a scenario or a user story needs further clarification, the development team may go back to the analysis activity. Otherwise, they proceed to the next step in the BDRE approach.
- Requirements validation and verification (V & V): This is the fourth step in the BDRE approach. The three Amigo's meetings take place again for refining the scenarios, produced from the modelling activity, according to their relevancy and importance. The purpose of this activity is to make sure that all the scenarios are done. To ensure that this happens, all the associated test cases of each scenario must successfully pass. Before producing the final set of scenarios, the development team negotiates and discusses all the examples with the customer's representative. In case of a disagreement, the three Amigos may decide to go back to the modelling activity or start over from the elicitation activity based on the severity level of the situation. Otherwise, the development team automates the produced final set of scenarios; thus, producing executable (aka. automated) specifications. The output of this V & V activity is the actual implementation, till the current development iteration, of the SUD.
- Requirements management: This is a cross-cutting activity in the BDRE approach through which all the other activities of the approach are maintained and managed.

V. Observational Study

This section presents an evaluation to investigate the feasibility and usefulness of the proposed BDRE approach.

A. Research Instruments

A research instrument is a tool that is used to measure, obtain, and analyse data subjects. Research instruments could be qualitative, quantitative, or a mix. In this observational study, a mixed approach seemed to be the better option as our level of understanding and familiarity with the product-under-study evolved throughout the lifetime of the development. The following are the research instruments [38] we used:

- Qualitative Methods: A qualitative research instrument is an exploratory tool that is used to have a better understanding of the subject at hand. It provides an in-depth look into the problem and/or helps developing ideas or solution hypotheses. In this research, we used two qualitative methods:
  - Observation: When using the observation research instrument, the observer can play the role of either a participant-observer or an observer participant. A participant-observer becomes a member of the community being observed; thus, enables them to earn the right to participate in the various activities accordingly. An observer participant, on the other hand, is treated as a visitor who can only observe the behaviour and the working environment of the development team, with no actual participation in their activities. Most of the time, we were an observer participant with few participations in some hands-on activities.

- Interviews: They are an integrated part of any agile-based environment. Interviews are basically a set of questions, regardless of their form (i.e., structured, semi-structured, unstructured, or a mixed-form interviews), with respective answers. Although agile advocates face-to-face communications, this might not be feasible at all times in practice. Alternatively, interviews can be mediated via telephones or other electronic means. We mainly used three types of interviews: in-depth interviews, face-to-face interviews, and discussion groups.

- Quantitative Methods: Quantitative research instruments are techniques that transform data from opinions/feelings into numbers and consequently from being subjective into being objective. One of the most popular quantitative research instruments is questionnaires. In this technique, questions can be in the format of multiple choices, dichotomous, short answers, checkboxes, drop-down, rating scales, and more. Depending on the research needs, one or more question formats can be adapted. In this research, we used the rating scale questions format. In this format, a participant is required to give an answer based on a well-defined evenly spaced range.

B. Working Environment

We tested the proposed approach in a small-sized (i.e., 100 – 200 employees) start-up agile-based company that is based in Egypt. The company has an intensive experience in agile development; in particular, Lean and Scrum agile methods.

The company focuses on the main agile practices such as iterative and incremental development; refactoring; automated testing; short iterations; pair programming; self-organising cross-functional teams; continuous deployment; progressive discovery; user story maps; and objectives and key results.

As the BDRE approach shares the same already implemented agile practices in place, the development team welcomingly embraced the proposed approach.

C. The Product under Development: RevoSuite

RevoSuite is a Business-to-Business Enterprise Software-as-a-service (SaaS). It is an artificial intelligence (AI)-enabled customer relationship management (CRM)/customer lifecycle management (CLM)/business intelligence (BI) system for pharmaceutical and life sciences businesses. The development of the product started in 2012 and evolved throughout the years. New enhancements are still added to the product despite being realised in the market late 2012.

D. The Observational Study Goal

The goal of this observational study is to investigate the feasibility of the BDRE approach in a real-life industrial case study. The elements of the observational study are inferred from the values of BDD. Table I lists the five elements of the observational study and the required observation from each one of them.

The participants in this study volunteered to take a part in our observational study. All the participants, except for the customer's representative, have worked on RevoSuite throughout its lifetime. The total number of volunteering participants is 24, categorised as follows: six business analysts, eleven developers, six quality assurance, and one customer's representative.

Prior to starting the observational study, we explained the BDRE approach to the participants and offered them training on how to implement the approach. Afterwards, the participants took parts in various complexity pilot projects throughout the RevoSuite different development iterations. Thus, enabled us to monitor and observe the participants' performance. Additionally, we developed a questionnaire addressing the elements listed in Table I in further details and asked our participants to anonymously answer the questionnaire from the perspective of each one's role.

TABLE I. OBSERVATIONAL STUDY ELEMENTS

Study Element	Required Observation
Learnability	Whether the participants are able to use the BDD ubiquitous language to express features, user stories, and scenarios
Coherence	Whether the participants are able to produce consistent outputs compared to that of the required business goals
Restrictions/Conflicts	Whether the participants are able to find all the explicit and implicit constraints and conflicts through executable specifications
Evolution	Whether the participants are able to start a feature, integrate new changes as they come in, and eventually deliver the feature in a manner consistent with the behaviour expected by the customer.
Readability	Whether the participants are able to read and understand the documentation, including the code, of the system.

VI. RESULTS AND DISCUSSION

This section presents and discusses the results of both the pilot projects and the questionnaire.

A. Pilot Projects Results

The participants' performance was measured by two factors: the time spent on each feature from beginning to end; and the uniformity of their output compared to that expected by the respective business goal. In general, the time spent on each feature was directly proportional to the complexity degree of that feature. Consequently, the time spent in high-complexity pilot projects varied between double to tribble that of the low-complexity projects. Despite that, the performance of all the participants was almost consistent regardless of the complexity of the features. The only exception was for the one customer's representative whose performance was inversely proportional to the complexity of the feature at hand.

In projects with low-medium complexity, we observed that:

- Learnability: almost all the participants were able to successfully use the BDD ubiquitous language to illustrate features, user stories, and scenarios.
- Coherence: more than 80% of the participants were able to have consistent outputs to those of the required business goals.
- Restrictions and conflicts: more than 75% of the participants were able to deduce all the explicit restrictions and conflicts. However, only half of them were able to spot all the implicit constraints.
- Evolution: more than 80% of the participants were able to start a feature, integrate new changes as they merge, and eventually deliver the feature (i.e., a core asset or a variability) in consistency with the expected behaviour of the system.
- Readability: all the participants were able to read and understand the system's documentation with minor difficulties.

In projects with high complexity, on the other hand, the participants spent more time on the features although they attained the same performance as that of the low-medium complexity projects. The only exception was the customer's representative whose performance dropped as the complexity of the feature increased.

B. Questionnaire Results

We used a five points Likert-scale, ranging from strongly disagree to strongly agree, to record the questionnaire responses. Fig. 1 illustrates the average responses per role for each question in the questionnaire. According to the recorded responses, the participants have come to a consensus that the BDRE approach is flexible, easy to understand, and easy to apply in practice. Some participants, however, shared their concerns about the potentiality and reliability of the BDRE approach in terms of scalability or when used with more complex systems. Lastly, finding implicit constraints was tricky and out of the comfort zone for some developers as well as for the customer's representative.

VII. Conclusion

APLE is increasingly gaining momentum in software development. Nonetheless, adopting APLE in practice calls for a special focus on RE. We proposed the BDRE approach to provide a flexible lightweight incremental RE process through using BDD throughout the different activities of RE. In this paper, we presented an observational study to examine five aspects of the BDRE approach in an empirical real case study. The results of the study were encouraging and shed the light on the strengths and weaknesses of the approach.

Acknowledgment

The authors would like to thank RevoSuite Company for their guidance and cooperation throughout the conduction of the observational study presented in this paper.

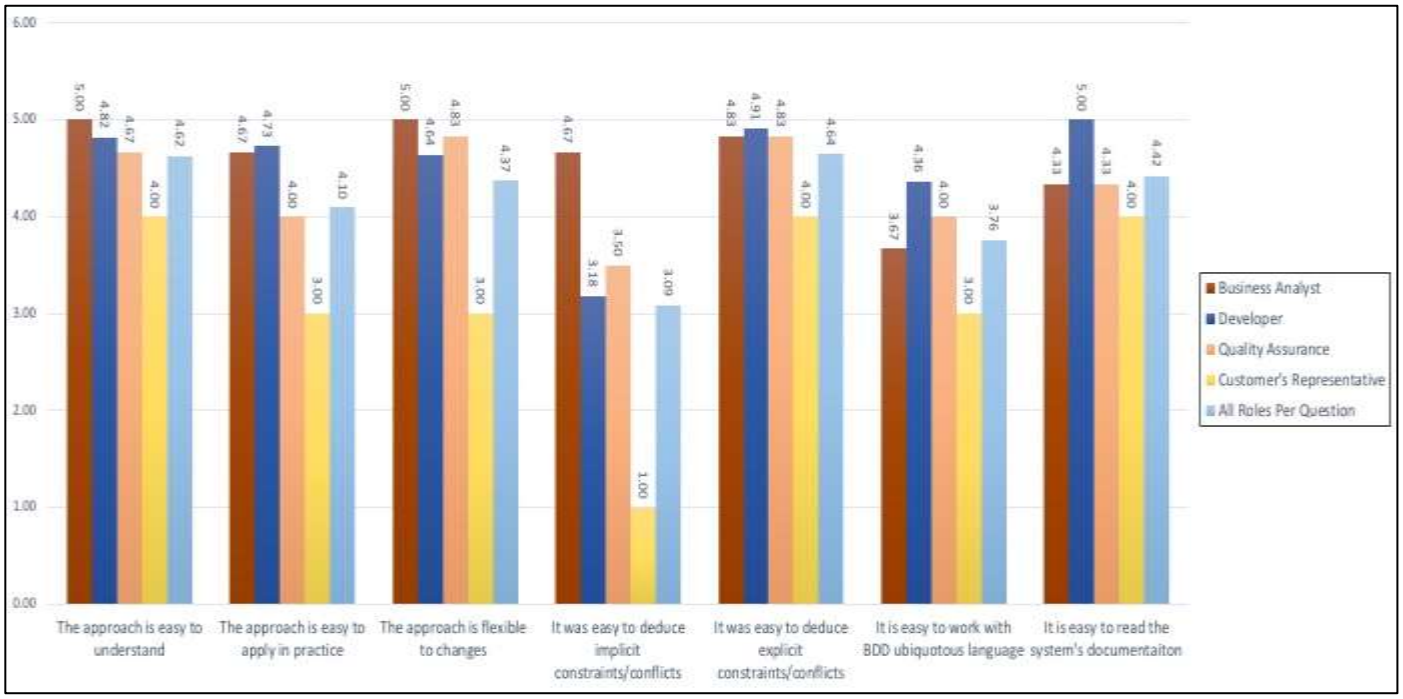


Fig. 1. Average Responses per Role to the Likert-Scale Questions.

REFERENCES

[1] K. Cooper, X. Franch, "APLE First International Workshop on Agile Product Line Engineering". IEEE Computer Society, pp. 205–206. Silver Spring, USA, 2006.

[2] K. Pohl, G. Böckle, F. Linden, Software Product Line Engineering: Foundations, Principles and Techniques. Springer, Germany 2005.

[3] L. Williams, A. Cockburn, Agile Software Development: It's About Feedback and Change, Computer 36(6), 39–43, 2003.

[4] D. North, Introducing BDD Available at: http://dannorth.net/introducing-bdd, 2006, last accessed 2018/1/1

[5] L. Westfall, Software Requirements Engineering: What, Why, Who, When and How?, Retrieved from http://www.westfallteam.com/Papers/The_Why_What_Who_When_and_How_Of_Software_Requirements.pdf, last accessed 2018/11/12

[6] Standish Group, The CHAOS Report 2015, Retrieved from https://www.standishgroup.com/sample_research, last accessed 2015/10/2

[7] G. Kontonya, I Sommerville, Requirements Engineering: Processes and Techniques. John Wiley & Sons, 1998.

[8] M. Chemuturi, Requirements Engineering and Management for Software Development Projects. Springer Publishing Company, 2012.

[9] D. Astels, A new look at test driven development. http://techblog.daveastels.com/files/BDD_Intro.pdf, last accessed 2019/3/1

[10] M.A. Noor, R. Rabiser, P. Grünbacher, "A collaborative approach for reengineering-based product line scoping", In: the 1st International Workshop on Agile Product Line Engineering (In conjunction with SPLC), 2006.

[11] M.A. Noor, R. Rabiser, P. Grünbacher, Agile Product Line Planning: A Collaborative Approach and a Case Study, Journal of Systems and Software, Vol. 81(6), pp. 868–882, 2007.

[12] M.A. Noor, P. Grünbacher, C. Hoyer, "A collaborative method for reuse potential assessment in reengineering-based product line adoption", In: Balancing Agility and Formalism in Software Engineering. LNCS, vol. 5082, pp. 69–83. Springer, Heidelberg, 2008.

[13] K. Feng, M. Lempert, Y. Tang, K. Tian, K. Cooper, X. Franch, "Developing a survey to collect expertise in agile product line requirements engineering", In: Agile 2007 Conference, International Research-in-Progress Workshop on Agile Software Engineering, pp. 1–4, 2007.

[14] K. Feng, Towards an Agile Product Line Requirements Engineering Framework: Knowledge Acquisition and Process Definition, Ph.D. Dissertation, The University of Texas at Dallas, 2009.

[15] P. Trinidad, D. Benavides, A. Durán, A. Ruiz-Cortés, M. Toro, Automated Error Analysis for the Agilization of Feature Modeling, Journal of Systems and Software, vol. 81(6), pp. 883–896, 2008 .

[16] R. Paige, X. Wang, Z. Stephenson, P. Brooke, "Towards an agile process for building software product lines", In: Proceedings of the 7th International Extreme Programming and Agile Processes in Software Engineering, Springer, Heidelberg, pp.198–199, 2006.

[17] G. Kakarontzas, I. Stamelos, P. Katsaros, "Product line variability with elastic components and test-driven development", In: Proceedings of the 2008 International Conference on Computational Intelligence for Modelling Control and Automation, IEEE Computer Society. 146–151, 2006.

[18] M. Raatikainen, K. Rautiainen, V. Myllärniemi, T. Männistö, "Integrating product family modeling with development management in agile methods", In:Proceedings of the 1st International Workshop on Software Development Governance, pp. 17–20, 2008.

[19] L. Taborda, The Release Matrix for Component-Based Software Systems, In: Proceedings of Component-Based Software Engineering, LNCS, vol. 3054, pp. 100–113, Springer, Heidelberg, 2004.

[20] R. Kurmann, "Agile SPL-SCM agile software product line configuration and release management", In: 1st International Workshop on Agile Product Line Engineering (In conjunction with SPLC), 2006.

[21] R. Carbon, M. Lindvall, D. Muthig, P. Costa, "Integrating product line engineering and agile methods: flexible design up-front VS. incremental design", In: Proceedings of the 1st International Workshop on Agile Product Line Engineering In conjunction with SPLC, pp. 1–8, 2006.

[22] R. Carbon, J. Knodel, D. Muthig, G. Meier, "Providing feedback from application to family engineering – The product line planning game at the Testo AG", In: Proceedings of the 12th International Software Product Line Conference, IEEE Computer Society, pp. 180–189, 2008.

[23] P. O'Leary, M.A. Babar, S. Thiel, I. Richardson, "Product derivation process and agile approaches: exploring the integration potential", In: Proceedings of the 2nd IFIP Central and East European Conference on Software Engineering Techniques. pp. 166–171, 2007.

[24] P. O'Leary, M.A. Babar, S. Thiel, I. Richardson, "Towards agile product derivation in software product line engineering", In: Proceedings of the 4th International Workshop on Rapid Integration of Software Engineering Techniques, pp. 19–32, 2007.

[25] P. O'Leary, S. Thiel, G. Botterweck, I. Richardson, "Towards a product derivation process framework", In: Proceedings of the 3rd IFIP TC2 Central and East European Conference on Software Engineering Techniques, pp. 189–202, 2008.

[26] P. O'Leary, F. McCaffery, I. Richardson, S. Thiel, "Towards agile product derivation in software product line engineering", In: Proceedings of the 16th European Conference on Software Process Improvement, pp. 81–86, 2009.

[27] P. O'Leary, R. Rabiser, I. Richardson, S. Thiel, "Important issues and aey Activities in product derivation: experiences from two independent research projects", In: Proceedings of the 13th International Software Product Line Conference, pp. 121–130, 2009.

[28] Y. Ghanam, S. Park, F. Maurer, "A test-driven approach to establishing and managing agile product lines", In: Proceedings of the 5th Software Product Lines Testing Workshop in conjunction with SPLC'08, pp. 151–156, 2008.

[29] Y. Ghanam, F. Maurer, "An iterative model for agile product line engineering", In: The SPLC Doctoral Symposium, 2008 - in conjunction with the SPLC'08, pp. 377–384, 2008.

[30] Y. Ghanam, F. Maurer, "Extreme product line engineering: managing variability and traceability via executable specifications", In: Agile Conference, AGILE '09, IEEE Computer Society, pp. 41–4, 2009.

[31] Y. Ghanam, F. Maurer, "Extreme product line engineering refactoring for variability: a test-driven approach", In: Proceedings of 11th International IV Conference on Agile Processes in Software Engineering and Extreme Programming, XP 2010, LNBIP, vol. 48, pp. 43–57, Springer, Heidelberg, 2010.

[32] F. F. Farahani, R. Ramsin, "Methodologies for agile product line engineering: a survey and evaluation", In: Proceedings of the 13th International Conference SoMeT_14, Amsterdam: IOS Press BV, pp.545-564, 2014.

[33] I.F. da Silva, P. Neto, P. O'Leary, E. de Almeida, S.R. de Lemos Meira, Agile Software Product Lines: A Systematic Mapping Study, Software: Practice and Experience, vol. 41(8), pp. 899–920, 2011.

[34] J. Díaz, J. Pérez, P.P. Alarcón, J. Garbajosa, Agile Product Line Engineering – A Systematic Literature Review, In: Software Practice and Experience, vol. 41(8), pp. 921–941, 2011.

[35] V. Alves, N. Niu, C. Alves, G. Valença, Requirements Engineering for Software Product Lines: A Systematic Literature Review, In: Information and Software Technology, vol. 52(8), pp. 806–820, 2010.

[36] D.F.S. Neiva, RiPLE-RE: A Requirements Engineering Process for Software Product Lines, M.Sc. Dissertation, Universidade Federal de Pernambuco, Brazil, 2009.

[37] H. Elshandidy, "Behaviour-driven requirements engineering for agile product line engineering", In: Proceedings of the 2019 IEEE 27th International Requirements Engineering Conference (RE), Jeju Island, Korea (South), pp. 434-439, 2019.

[38] R. Trigueros, M. Juan, F. Sandoval, Qualitative and Quantitative Research Instruments Research tools, 2017.

Detecting Generic Network Intrusion Attacks using Tree-based Machine Learning Methods

Yazan Ahmad Alsariera
Department of Computer Science, Faculty of Science
Northern Border University, Arar 73222, Kingdom of Saudi Arabia

***Abstract*—The development Intrusion Detection System (IDS) has a solid impact in mitigating against internal and external cyber threats among other cybersecurity methods. The machine learning-based method for IDS has proven to be an effective approach to detecting either anomaly or multiple classes of intrusion. For the detection of various types of intrusion by a single IDS model, it is discovered that the overall high accuracy of the IDS model does not translate to high accuracy for each attack type. Some intrusion attacks are seen to share similarities with other attacks thereby evading detection, one of which is the generic attack. The notoriety of the generic attack is the ability of a single generic attack to compromise a whole bunch of block-ciphers. Therefore, this study proposed a machine learning framework to specifically detect generic network intrusion by implementing two (2) decision tree algorithms. The decision tree methods were developed using two distinct variants namely the J48 and Random Tree algorithms. A balanced generic network dataset was curated and used for model development. A 10-fold cross-validation technique was implemented for model development and performance evaluation, where all obtainable performance scores were extracted and presented. The performances of the decision tree methods for generic network intrusion attack detection were comparative analysis and also evaluated against existing methods. The proposed methods of this study are robust, stable and empirically seen to have outperformed existing methods.**

Keywords*—Generic attack; decision trees; cybersecurity; intrusion detection

I. Introduction

The unprecedented surge of digital users over the years had led to the expansion of the world's cyberspace [1], [2]. Technological advancements had seen the enablement and rapid growth of various digital services offered to individuals and entities across the world [3]. Cyberspace consists of billions of connected devices and users whose security is now pivotal to the existence of the modern world [4], [5]. Cybersecurity emerges as the field that ensures the security of cyberspace.

Cybersecurity ensures data, information, and devices confidentiality, availability, and integrity against cyberspace attackers through sets of systems, technologies, and processes [6]. That means cybersecurity is responsible for providing countermeasures for removing and or ameliorating security threats and breaches (internal or external intrusion attacks) [7].

Before the execution of any known and unknown threat or attack, an attacker must first intrude (i.e. gain access to) his or her target network. This made the detection of intrusion a pivotal research area in cybersecurity [8]. The development of Intrusion Detection Systems (IDS) has received enormous research spotlight and the application of machine learning algorithms has proven to be the best method of developing effective and efficient IDS among other methods [9] – [13].

A recent review of the literature identified a problem that the effectiveness of machine learning (ML) based IDS for classifying multiple types of intrusion using a single model are hampered among network attacks with similar characteristics [14]. Hence, it becomes necessary to isolate and develop specific machine learning IDS for these types of extremely dangerous attacks. One such dangerous attack is called the 'Generic' attack. The generic attack is dangerous such that one (1) generic attack can attack all block-ciphers regardless of the distinct structure of the ciphers [15], [16].

Despite known to be dangerous, countermeasures against generic intrusion are not well researched and developed in the context of applying ML algorithms. Generic network intrusions are not captured by KDDCup'99 and NSL-KDD intrusion network dataset [1]. However, the comprehensive and contemporary dataset published by [16] contains generic attack traffics. Even so, this dataset [16] is usually used for developing anomaly (i.e. normal and attack) [17] and multi-class (i.e. normal and nine (9) other attacks) IDS [18]. Hence, this study is motivated and thereby proposes an ML-based IDS framework specifically for detecting generic network intrusion attack. The contributions to knowledge made by this research are highlighted below:

1) Development of a balanced network intrusion 'Generic' attack dataset for machine learning classification process; and

2) Implementation and performance evaluation of two (2) distinct machine learning decision tree algorithms as the proposed methods for detecting generic network intrusion.

The decision trees were selected as they are seen to have a sharp distinction between their methods of learning, unlike other decision tree algorithms. More so, through this research work, answers to the following research questions are sought:

1) How well can J48 and Random Tree decision tree algorithms effectively detect generic attacks?

2) Is there any significant difference(s) in using distinct variants of decision tree algorithms for detecting generic network intrusion?

3) How good is the performance of the proposed method against related existing methods?

The proposed methods of this study, whose application lies in network security, will serve as a customized IDS for detecting generic network intrusion. The remaining part of this study is structured as follow: Section II contains a review of related works, Section II vividly reveals the method (i.e. dataset, implemented models and performance evaluation metrics), Section IV present results and discussions and finally Section V shares the conclusion and future works.

II. Review of Related Works

Although stand-alone researches on generic network intrusion detecting are very scarce, some multi-classification researches on IDS present a breakdown of their model's performances. The type of research studies that made these provisions as well as other closely related studies was sought and reviewed accordingly.

The research work of [19] presented an ensemble of sophisticated deep learning algorithms for detecting different types of network anomalies. The study implemented a majority voting ensemble of three hyper-parameter long-short-term memory deep neural network with an embedded feature extraction module. The feature extraction module composed of a Genetic Algorithm (GA). This algorithmic framework was implemented and fitted on NSL-KDD and UNSW-NB15 datasets. The published method reported an overall accuracy of 99.9%. However, the performance of the implemented framework for the detection of a generic attack dropped to 95.23% without feature selection and 97.31% with feature selection. This supports the need to develop a specified generic network IDS method with increased accuracy and lower false alarm rate.

Another study [15], presented a novel integrated rule-based multi-classification method IDS fitted on the UNSW-NB15 dataset. The proposed method is a misuse-based IDS for four type of attacks namely: DOS, Generic, Exploit, Probe and the Normal traffic in a network. The proposed method achieved an overall Average accuracy (i.e. AvgAcc) of 65.21% for all classes of attacks and a False Alarm Rate of 2.01%. From the study, an improved IDS is generally required even to detect other types of network intrusion.

A more recent study [20] published a stacked ensemble method for developing a multi-classifying IDS. Three (3) methods for stacking base classification algorithms were implemented namely: Meta Decision Tree (MDT), Multiple Model Trees (MMT) and Multi-Response Linear Regression (MLR). The base classifiers are Naïve Bayes, Decision Tree and K-Nearest Neighbour. The evaluation of the base learner (DT) for classifying all attack type achieved an overall accuracy of 75.71% without feature selection. The MMT ensemble method produced 96.89% overall accuracy, the MDT ensemble method had 98.08% and the MLR method had 97.8% accuracies based on the correlated reduced feature selected model. The performance of the method for detecting generic network intrusion was not disclosed.

Gharaee & Hosseinvand [21] reportedly developed a new feature selection IDS using the support vector machine algorithm and a genetic algorithm for feature selection which was referred to as "GF-SVM". The genetic feature selection algorithm was reportedly developed using a novel fitness function that was responsible for dimensionality reduction. The overall performance of the multi-classifier IDS was broken down and presented for each class of attack. The implemented method was able to achieve an accuracy of 97.51%, 96.69% True Positive (TP) rate, and 0.01% False Positive (FP) rate for the 'Generic' attack as related to this study. The rate at which generic network intrusion can be detected (TP) by [21] can be further improved while the FP rate can be lowered which is the intention of this study.

Succinctly, the review of related literature that provides the performance breakdown of the existing method further strengthen the need for developing stand-alone generic network intrusion attack detectors.

III. Method

In this section, details of the dataset of the study are presented as well as the machine learning algorithms used to implement the generic detector IDS and the performance evaluation metrics for the implemented models.

A. Dataset

Dataset serves as a core part of empirical research. Therefore, it is important to make use of the dataset that truly serves the study's aim, strengthens the study as well as being state-art-of-the-art. In this study, the development of ML decision trees methods for detecting 'generic' network intrusion attack is crucial. Therefore, a state-of-the-art dataset is used to conduct the study's experiment. In the research scope of developing ML methods for network intrusion detection, the UNSW-NB15 dataset is currently the best benchmarking and the openly available dataset [16]. This dataset ousted other public network attack datasets (i.e. KDDCup'99 and NSL-KDD) by providing contemporary network traffic and attacks [16].

The KDDCup'99 is the initial benchmarking dataset but was revised and led to the production of the NSL-KDD dataset. The NSL-KDD data is devoid of all redundancy in its predecessor and provides a more balanced dataset [9], [22]. However, it does not contain contemporary attacks as executed by attackers' such as the 'Generic' attack type that is been studied in this research work. More so, attackers daily carry out dynamic attacks which then require developing intelligent countermeasures from a contemporary dataset (having real and or synthetic attacks) to adequately ameliorate novel malicious network activities [23]. KDDCup'99 and NSL-KDD are not reflective of contemporary attacks and network packets, which single-out and justifies the usage of the UNSW-NB15 dataset by this study.

As mentioned in the introduction section, a high-performing multi-classifier does not usually achieve single-class discrimination when two or more class shares similar feature values [14]. Therefore, to achieve the aim of this study, the 'Generic' attack instances were extracted from the UNSW-NB15 dataset alongside adequate 'Normal' instances to create a balanced dataset. Table I gives insights into the dataset used in this study.

The original dataset contains forty-five (45) variables which were reduced to forty-three (43) in the newly created dataset (i.e. without the 'id' and 'attack_cat' variables). Forty-two (42) of all variables serves as independent variables and the forty-third (43rd) variable is the dependent variable with two values as shown in Table I. From the original dataset, there are 18,871 'Generic' attack instances. As such, 18,954 normal instances were extracted to create the benchmark dataset for this study.

B. Implemented Models

In this study, two (2) distinct variants of the decision tree (DT) machine learning algorithms were used to fit the required models for detecting generic network intrusion attacks.

Decision tree algorithms are a family of machine learning classification and regression algorithms that fits a model on a given dataset having considered the entropy of some or all attributes for making its splitting decision. Tree-based machine learning algorithms are widely used and acceptable for various research and industrial areas, even as distant as software defect prediction in the field of software engineering [24] and even for the prediction of factors in educational management [25]. Decision Tree models are known to always produce interpretable models. Additionally, the derived tree inherent in every decision tree model can be used as a rule(s) for guiding expert decision aside from its usage for prediction.

Fundamentally, all decision tree algorithm can perform both regression and classification (primarily binary classification) analyses. Decision algorithms usually fit its model through a greedy top-down method which is performed recursively on the dataset to find the most informative variable at each split decision junction [25]. Additionally, it may also include a method for producing a fine-tuned tree by the way of pruning the initial tree based on the error rate thereby removing redundant branches [26]. All decision tree algorithms begin the process of fitting a model with a root node (which is the most informative variable) and then create branches and some leaves downwardly based on the results of testing variables values Extracted from [26].

Pseudocode 1: A typical Decision Tree Algorithm.

```
1:   Create a root node R;
2:   IF (W belongs to same category C)
       {leaf node = R;
         Mark R as class C;
         Return R;
       }
3:   For i=1 to R
       {Calculate Information_gain (Ai);}
4:   ta = testing attribute;
5:   R.ta = attribute having highest information_gain;
6:   If(R.ta == continuous)
       {find threshold;}
7:   For (Each W in splitting of W)
8:     If (W is empty)
         {child of R is leaf node;}
         else
         {child of R= dtree W;}
9:   calculate the classification error rate of node R
10:  return R;
```

TABLE I. TABLE TYPE STYLES

Dataset Description		
No of Attributes		43
No. of Independent Variables		42
Dependent Feature values distribution	***Values***	
	Generic	***Normal***
	18,871	18,954

As mentioned earlier, two (2) machine learning decision trees algorithms variants were considered in this study. These are the famous J48 and Random Tree algorithms.

J48 algorithm is usually a greedy top-down approach starting from the root node through the branches down to the leaves. It can also follow a bottom-up approach. It contains decision nodes (branches) which are indicators to tested attributes and leaves which signifies class values. J48 is characterized by its ability to accept both nominal and continuous variable values. Also, it includes an imputation technique that resolves missing values in variables as well as a pruning mechanism for developing optimal but small trees that avoid over-fitting [26]. In this study, the J48 algorithm was implemented and fitted on the described dataset. The resulting model was evaluated using all obtainable performance evaluation metrics.

On the other hand, Random Tree is another variant of the decision tree algorithm family that fits various decisions trees on a given dataset using N randomly selected variables at each node. These sets of random decision trees usually form a uniform distribution which gives each tree an equal sampling chance. These uniformly distributed trees are used to develop a random tree through aggregation which produce a more robust and accurate model. In this study, the Random Tree algorithm was implemented and fitted on the dataset producing a model which was subjected to the evaluation of its performance in discriminating between 'Generic' intrusion attack and normal network traffic.

The experimental framework of this study is graphically depicted in Fig. 1 which illustrate how the data preprocessing and processing, the selected machine learning decision trees methods were developed and their respective performance evaluation.

The decision tree methods, namely J48 and Random tree decision tree algorithms, were implemented and fitted on the randomly shuffled dataset through the 10-fold cross-validation technique. The cross-validation technique is the method of fitting a robust model by splitting the dataset into user-defined value – 10 partitions. It trains the model using the first 9 splits and test on the set-aside split. This is repeated 10 times until all splits are used for training and testing. The 10 models are then aggregated to produce a robust model. The performances of the fitted models (i.e. J48 and Random Tree generic attack detectors) were measure and evaluated using widely acceptable metrics, such as confusion matrix, MCC, accuracy, True positive, True negative, kappa score and others as previously mentioned.

C. Performance Evaluation Metrics

This section discusses how the performance of the proposed ML decision trees methods for detecting generic network intrusion attack was evaluated. The models can be referred to as binary classification (i.e. two class values) methods. Thus, our proposed methods were evaluated by populating and reporting their respective performance values using the confusion matrix. More so, other performance values, which can be derived from the confusion matrix, were also reported. These are: TP Rate (i.e. Detection Rate), FP Rate (i.e. False Alarm Rate), Precision, Recall, F-Measure, Matthews Correlation Coefficient (MCC), Area Under Curve (AUC) [2], [10], [19], [27]–[30]. Also, the overall accuracy (i.e. the percentage of correctly classified 'Generic' attack and normal network traffic), as well as kappa value, were obtained for each method.

For emphasis, the MCC metric is arguably the prime metric for evaluating a binary classification as it based on all four values of the confusion matrix [19], [31]. It reveals the correlation coefficient among the detected and expected predictions, having a value ranging from 0 to 1 [30]. Therefore, a better gauge of the classification model is revealed in the MCC value. However, this does not relegate other performance metrics. MCC metric is calculated as seen in Equation 1.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

Matthews Correlation Coefficient extracted from [31].

As illustrated in the proposed empirical framework presented in Fig. 1, all 'Generic' attack network instances from the UNSW-NB15 dataset were extracted. Additionally, enough normal network instances were also extracted to create a balanced dataset that serves as input to the decision tree methods. Before inputting the balanced dataset, it was shuffled to ensure instances of both class values were properly mixed and the model can learn from the distribution simultaneously.

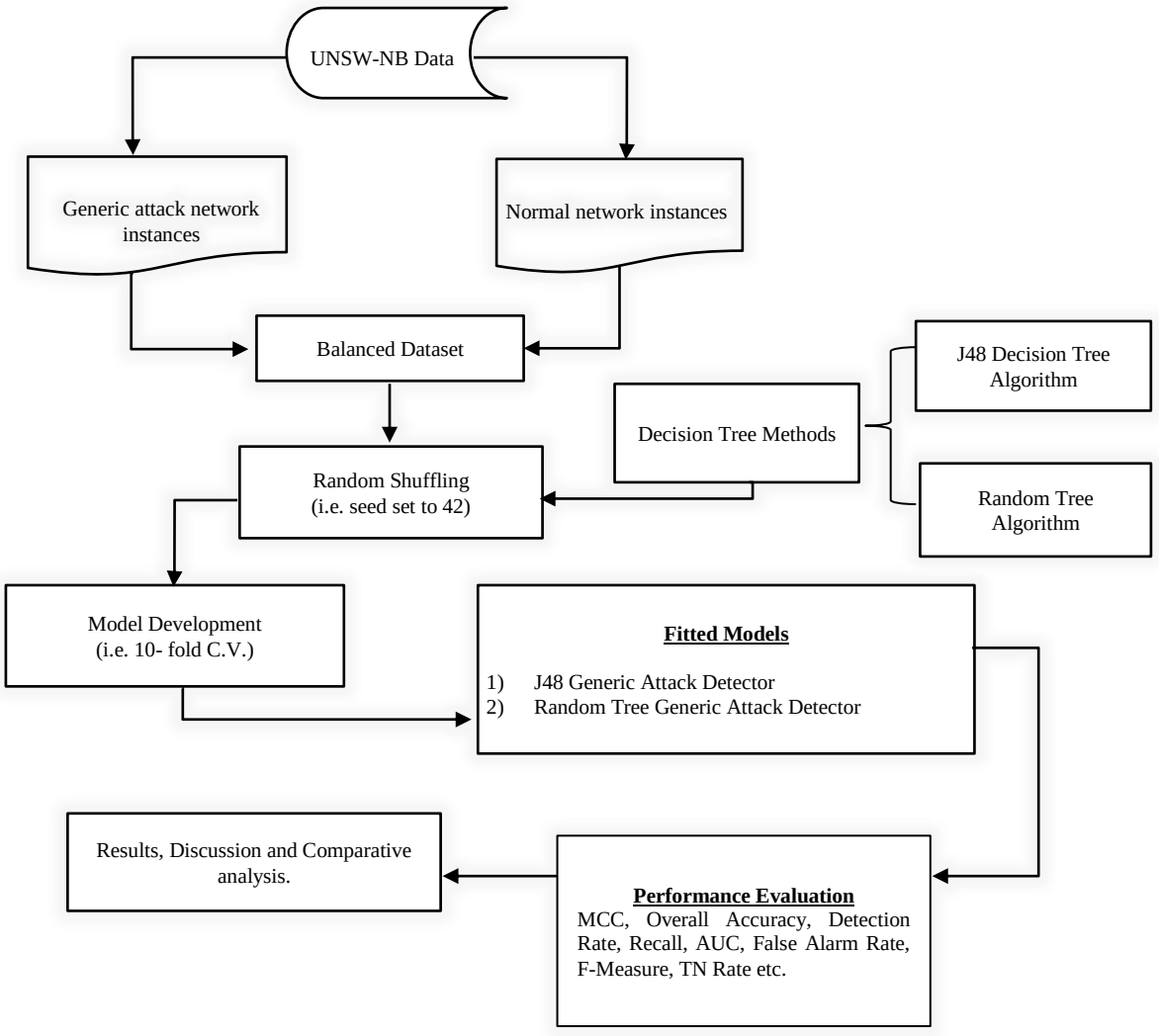


Fig. 1. Experimental Framework.

IV. Results and Discussion

This section reports the performance of the proposed ML decision trees methods in tables and charts. More so, the results are being discussed individually and a comparative analysis of the performance of the proposed methods against existing methods is provided by answering the aforementioned research questions.

A. Results

As previously discussed, the dataset has a total number of 18,871 generic instances and 18,954 normal instances. Also, the proposed methods were developed using a 10-fold cross-validation technique for classification model development.

The J48 'Generic' attack classifier performance confusion matrix is presented in Table II and other derived performance evaluation values are presented in Table III.

As seen in Table II, there were 18,830 correctly classified instances of 'generic' attack out of the total of 18,871. Similarly, there were 18,921 correctly classified normal instances out of the total of 18,954. A total of 41 generic attack instances were misclassified as normal while a total of 33 normal instances were falsely classified as generic.

The values of other derived performance measures are revealed in Table III.

From Table III, the proposed J48 classifier achieved an overall accuracy of 99.804% - an excellent performance. This is evident as revealed in the confusion matrix in Table II. Additionally, this model scores a kappa value of 0.9961 indicate the model performed higher than chance. The model achieved the TP and TN rates of 0.998 respectively while it also had FP and FN rates of 0.002 respectively. Its precision, f-measure and recall values also tallied at 0.998. Lastly, it had an AUC score of 0.999 while it had an MCC score of 0.996.

TABLE II. J48 Generic Attack Discriminator Confusion Matrix

	Generic	***Normal***
Generic	18,830	41
Normal	33	18,921

TABLE III. J48 Generic Attack Discriminator Evaluation

Evaluation Metric	**J48's Performance Value**
Accuracy	99.804%
Kappa	0.9961
TP Rate (Detection Rate)	0.998
FP Rate	0.002
TN Rate	0.998
FN Rate	0.002
Precision	0.998
Recall	0.998
F-measure	0.998
MCC	0.996
AUC	0.999

The performance of the model obtained after fitting the Random Tree algorithm on the dataset via 10-fold cross-validation was also evaluated just like its counterpart. Table IV presents the confusion matrix for the Random tree classifier.

From Table IV, 18,776 of 18,871 generic attack instances were correctly classified while 18,883 of 18,954 normal instances were also correctly classified. 95 generic instances were misclassified as normal traffic while 71 normal instances were misclassified as a generic attack.

Additionally, other performance values were derived and depicted in Table V.

This Random tree proposed method achieved an overall accuracy of 99.561% and a kappa score of 0.9912. It obtained a TP rate of 0.995, TN rate of 0.996, FP Rate of 0.004, and FN Rate of 0.005. More so, it had a precision, recall and f-measure score tallied at 0.996 respectively. The classifier scored an AUC value of 0.997 while having a 0.991 MCC score.

B. Discussion

This study aims to develop a machine learning framework specifically capable of detecting the extremely dangerous generic network intrusion attack which shares similarity with other types of attack thereby evades detection. Following the implemented of the proposed framework, the two generic network intrusion attack detectors were robustly developed and evaluated using the 10-fold cross-validation technique. Two algorithmically distinct decision tree methods were developed, and all obtainable performance evaluation scores were derived from the confusion matrix obtained produced by each of the methods.

TABLE IV. Random Tree Generic Attack Discriminator Confusion Matrix

	Generic	***Normal***
Generic	18,776	95
Normal	71	18,883

TABLE V. Random Tree Generic Attack Discriminator Evaluation

Evaluation Metric	**RT's Performance Value**
Accuracy	99.561%
Kappa	0.9912
TP Rate (Detection Rate)	0.995
FP Rate	0.004
TN Rate	0.996
FN Rate	0.005
Precision	0.996
Recall	0.996
F-measure	0.996
MCC	0.991
AUC	0.997

A comparative analysis of both proposed methods empirical results reveals that the model produced after fitting the J48 decision tree algorithm is insignificantly better than Random Tree's model as presented in Table VI.

Table VI present the empirical results of the methods using four benchmark performance metrics out of all performance metrics mentioned in the performance evaluation section. The J48 DT method is seen to produce an overall accuracy of 99.8% while the Random Tree DT method produced an overall accuracy of 99.6%. Similar trends are recorded for the detection rate (i.e. TP) and the False Alarm Rate (i.e. FP). J48 method detected a 'Generic' attack at 99.8% while Random Tree did the same at 99.5%.

More so, both methods were able to detect generic network intrusion attack with an extremely low false alarm rate. J48% false alarm rate is 0.002% while the Random Tree method's rate is at 0.004%. Both decision tree methods proved to be a viable method for detecting a generic attack. This summarized comparative analysis is depicted in Fig. 2.

C. Comparative Analysis with Existing Methods

The answers to this study's research questions, which also facilitate comparative analysis of the proposed methods even with existing methods, are provided in this section. The first research question is about the effectiveness of the machine learning decision tree (i.e. J48 and Random Tree) methods. As seen through the empirical results, the effectiveness of these methods for detecting generic network cannot be over-emphasized. Both methods are excellently effective at a detection rate not lower than 99% and with an incredible false alarm rate lower than 0.05%. The MCC scores of both methods were also not lower than 0.99 as well as their precision, recall, f-measure and ROC values. All these results indicate that both J48 and Random Tree generic network intrusion attack detector are highly effective.

TABLE VI. EMPIRICAL RESULTS

Decision Tree Algorithms	*Accuracy (%)*	*Detection Rate (%)*	*False Alarm Rate*	*MCC*
J48	99.8	99.8	0.002	0.996
Random Tree	99.6	99.5	0.004	0.991

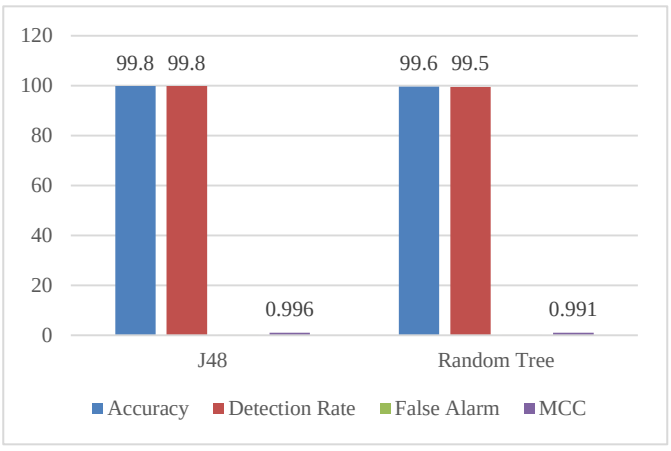


Fig. 2. Summative Comparative Analysis of Proposed Methods.

The second research question aims to investigate the comparative performance of the implemented distinct variants of the decision trees method. Also, the empirical results indicate that the methods do not significantly outperform each other. Since all other variants of the decision tree algorithms are closely related to one of these distinct variants, it is safe to infer that any decision tree method implemented for detecting generic attack will perform similarly to the proposed methods of this study.

Lastly, the answer to the third research question which is the comparative analysis of the proposed method against the existing method is provided. The published sophisticated genetic algorithm and deep-learning method [19] reported a 95.23% overall accuracy for detecting generic attack with feature selection. This reported performance [19] which is lower compared to the overall accuracy for the said method for multi-classification, is also lower than the performance of this study's proposed methods for generic network intrusion detection.

Also, the decision tree method published by [20] achieved a 75.71% generic attack detection accuracy without feature selection while its stacked ensemble methods on the correlation reduced models produced 97.8%, 96.89% and 98.08% accuracies. All these models were outperformed by the proposed methods of this study as this study's methods had at least a 99% detection rate.

Additionally, the novel integrated rule-based IDS [15] for detecting DOS, Generic, Exploit, Probe attacks and the Normal traffic in a network had an overall AvgAcc of 65.21% for all classes of attacks and a False Alarm Rate of 2.01% which is comparatively lower than the performance of this study's method even if broken down into different attack types. The existing (i.e. multi-classification) methods are low-performing machine learning methods for detecting generic network intrusion attack which justifies the importance of this research.

V. CONCLUSION AND FUTURE WORKS

This study proceeds to develop tree-based machine learning generic network intrusion detection models, having identified the problem that generic attack shares similarities with other attacks and usually evades detection from multi-classification IDS. Two (2) distinct tree-based machine learning method, J48 and Random Tree algorithms were proposed to implement this study's models.

J48 model was able to detect generic network attack at 99.8% and a false alarm rate of 0.002 while the Random Tree model detected generic network attack at a 99.6% detection rate and a false alarm rate of 0.004. The comparative analysis of the proposed methods against existing methods which are mostly multi-classification IDS reveals that the proposed method performed better than all of them in detecting generic network intrusion.

In the future, the application of other types or families of machine learning classification method will be explored. More so, the culling out of important feature (i.e. reducing the dimensionality) from the original feature space of this balanced generic attack dataset will be considered.

REFERENCES

[1] Y. Xin et al., "Machine Learning and Deep Learning Methods for Cybersecurity," IEEE Access, vol. 6, pp. 35365–35381, 2018.

[2] O. A. Sarumi, A. O. Adetunmbi, and F. A. Adetoye, "Discovering computer networks intrusion using data analytics and machine intelligence," Sci. African, vol. 9, p. e00500, 2020.

[3] H. N. Thanh and T. Van Lang, "Evaluating Effectiveness of Ensemble Classifiers When Detecting Fuzzers Attacks on the Unsw-Nb15 Dataset," J. Comput. Sci. Cybern., vol. 36, no. 2, pp. 173–185, 2020.

[4] P. Kumar, G. P. Gupta, and R. Tripathi, "A distributed ensemble design based intrusion detection system using fog computing to protect the internet of things networks," J. Ambient Intell. Humaniz. Comput., no. 0123456789, 2020.

[5] O. Faker and E. Dogdu, "Intrusion detection using big data and deep learning techniques," ACMSE 2019 - Proc. 2019 ACM Southeast Conf., pp. 86–93, 2019.

[6] M. R. Gauthama Raman et al., An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm, vol. 53, no. 5. Springer Netherlands, 2020.

[7] V. Dutta, M. Choraś, R. Kozik, and M. Pawlicki, "Hybrid model for improving the classification effectiveness of network intrusion detection," in Complex, Intelligent, and Software Intensive Systems, 2020, vol. Springer, pp. 405–414.

[8] W. Wei, S. Chen, Q. Lin, J. Ji, and J. Chen, "A multi-objective immune algorithm for intrusion feature selection," Appl. Soft Comput. J., vol. 95, p. 106522, 2020.

[9] M. A. Mabayoje, A. O. Balogun, A. O. Ameen, and V. E. Adeyemo, "Influence of Feature Selection on Multi-Layer Perceptron Classifier for Intrusion Detection System," Comput. Inf. Syst. Dev. Informatics Allied Res. J., vol. 7, no. 4, pp. 87–94, 2016.

[10] A. V. Elijah, A. Abdullah, N. Z. JhanJhi, M. Supramaniam, and A. O. Balogun, "Ensemble and deep-learning methods for two-class and multi-attack anomaly intrusion detection: An empirical study," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 9, pp. 520–528, 2019.

[11] A. O. Balogun, A. M. Balogun, V. E. Adeyemo, and P. O. Sadiku, "A Network Intrusion Detection System : Enhanced Classification via Clustering," Comput. Inf. Syst. Dev. Informatics Allied Res. J., vol. 6, no. 4, pp. 53–58, 2015.

[12] P. Illy, G. Kaddoum, C. M. Moreira, K. Kaur, and S. Garg, "Securing Fog-to-Things Environment Using Intrusion Detection System Based On Ensemble Learning," no. April, pp. 15–18, 2019.

[13] M. Idhammad, K. Afdel, and M. Belouch, "Semi-supervised machine learning approach for DDoS detection," Appl. Intell., vol. 48, no. 10, pp. 3193–3208, 2018.

[14] T. Salman, D. Bhamare, A. Erbad, R. Jain, and M. Samaka, "Machine Learning for Anomaly Detection and Categorization in Multi-Cloud Environments," Proc. - 4th IEEE Int. Conf. Cyber Secur. Cloud Comput. CSCloud 2017 3rd IEEE Int. Conf. Scalable Smart Cloud, SSC 2017, pp. 97–103, 2017.

[15] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, "An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset," Cluster Comput., vol. 23, no. 2, pp. 1397–1418, 2020.

[16] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in 2015 Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings IEEE, 2015, pp. 1–6.

[17] F. Feng, X. Liu, B. Yong, R. Zhou, and Q. Zhou, "Anomaly detection in ad-hoc networks based on deep learning model: A plug and play device," Ad Hoc Networks, vol. 84, pp. 82–89, 2019.

[18] M. Nawir, A. Amir, N. Yaakob, and O. N. G. B. I. Lynn, "Multi-Classification of Unsw-Nb15 Dataset for Network Anomaly Detection System," J. Theor. Appl. Inf. Technol., vol. 96, no. 15, pp. 5094–5104, 2018.

[19] I. S. Thaseen, A. K. Chitturi, F. Al‐Turjman, A. Shankar, M. R. Ghalib, and K. Abhishek, "An intelligent ensemble of long‐short‐term memory with genetic algorithm for network anomaly identification," Trans. Emerg. Telecommun. Technol., no. September, pp. 1–21, 2020.

[20] O. O. Olasehinde, "A Stacked Ensemble Intrusion Detection Approach for the Protection of Information System," Int. J. Infomation Secur. Res., vol. 10, no. 1, pp. 910–923, 2020.

[21] H. Gharaee and H. Hosseinvand, "A new feature selection IDS based on genetic algorithm and SVM," 2016 8th Int. Symp. Telecommun. IST 2016, pp. 139–144, 2017.

[22] A. Salih, X. Ma, and E. Peytchev, "Detection and Classification of Covert Channels in IPv6 Using Enhanced Machine Learning," 2015.

[23] G. Li and Z. Yan, "Data Fusion for Network Intrusion Detection : A Review," vol. 2018, 2018.

[24] A. O. Balogun et al., "SMOTE-Based Homogeneous Ensemble Methods for Software Defect Prediction," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12254 LNCS, pp. 615–631, 2020.

[25] W. Gata et al., "Prediction of Teachers' Lateness Factors Coming to School Using C4.5, Random Tree, Random Forest Algorithm," vol. 258, no. Icream 2018, pp. 161–166, 2019.

[26] S. Aljawarneh, M. B. Yassein, and M. Aljundi, "An enhanced J48 classification algorithm for the anomaly intrusion detection systems," Cluster Comput., vol. 22, no. 5, pp. 10549–10565, 2019.

[27] S. M. Kasongo and Y. Sun, "A deep learning method with wrapper based feature extraction for wireless intrusion detection system," Comput. Secur., vol. 92, 2020.

[28] Y. A. Alsariera, A. V. Elijah, and A. O. Balogun, "Phishing Website Detection: Forest by Penalizing Attributes Algorithm and Its Enhanced Variations," Arab. J. Sci. Eng., vol. 45, no. 12, pp. 10459–10470, 2020.

[29] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, A. K. Alazzawi, "AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites," IEEE Access, vol. 8, pp. 142532-142542, 2020.

[30] J. O. Mebawondu, O. D. Alowolodu, J. O. Mebawondu, and A. O. Adetunmbi, "Network intrusion detection system using supervised learning paradigm," Sci. African, vol. 9, 2020.

[31] N. Li, M. Shepperd, and Y. Guo, "A Systematic Review of Unsupervised Learning Techniques for Software Defect Prediction," 2019.

An Extensive Analysis of the Vision-based Deep Learning Techniques for Action Recognition

Manasa R[1], Ritika Shukla[2], Saranya KC[3]
School of Electronics Engineering
Vellore Institute of Technology
Vellore, India

***Abstract*—Action recognition involves the idea of localizing and classifying actions in a video over a sequence of frames. It can be thought of as an image classification task extended temporally. The information obtained over the multitude of frames is aggregated to comprehend the action classification output. Applications of action recognition systems range from assistance for healthcare systems to human-machine interaction. Action recognition has proven to be a challenging task as it poses many impediments including high computation cost, capturing extended context, designing complex architectures, and lack of benchmark datasets. Increasing the efficiency of algorithms in human action recognition can significantly improve the probability of implementing it in real-world scenarios. This paper has summarized the evolution of various action localization, classification, and detection algorithms applied to data from vision-based sensors. We have also reviewed the datasets that have been used for the action classification, localization, and detection process. We have further explored the areas of action classification, temporal and spatiotemporal action detection, which use convolution neural networks, recurrent neural networks, or a combination of both.**

Keywords*—Action recognition; deep learning; vision sensors; convolution neural networks (CNN); recurrent neural networks (RNN); action classification; temporal action detection; spatiotemporal action detection*

I. Introduction

There are two types of human action recognition systems - sensor-based and video-based [1]. Various on-body and ambient sensors are used to understand and label human actions performed in recorded videos or real-time video streaming. Video cameras are the essential wellsprings of new data on the Internet. A video is an organized arrangement of frames of a similar resolution taken at regular intervals of time. While developing the video processing algorithm, the video is partitioned into two classes-video streams and video sequences. Video stream is a continuous video for online processing as we are unaware of the information present in future frames. The video sequence is a fixed-length video where all frames are accessible without a moment's delay. Currently, most video cameras do not perform automated action recognition. Since the amount of video data available is extremely high, automatic action recognition has become a necessity. Furthermore, action recognition will facilitate efficient human-machine interactions, video surveillance, patient-care, smart homes, sports video analysis, gaming, and intelligent retail.

An action recognition process involves two tasks: action classification and action localization, as represented in Fig. 1. Action classification consists of assigning labels to various action instances in videos. Although it is possible to classify some actions using single frames, most actions occur in a series of adjacent frames. The motion in these frames must be captured to classify the actions. Video data brings a new feature that is absent in static images, which are motion. This motion characterizes actions in videos. To obtain these motion features, the motion field must be obtained. Optical flow, which represents the apparent motion between frames, is used to estimate the motion field.

The extensive input data, less availability of computational resources, and difficulty in obtaining the optical flow pose major problems while classifying actions. In action classification tasks, the model must run through multiple windows in search of action instances. This is computationally expensive and time-consuming. Temporal action detection models work on the data before action classification models to reduce computational costs. They define the temporal bounds of action instances and specify to the action classification model the actions' temporal location in any given video sequence. Spatiotemporal action detection models provide information on the spatial locations of the action in addition to the temporal bounds.

The field of computer vision and deep learning has already seen significant success in object detection, classification, and localization techniques, and now the area of study is moving towards efficient action detection and recognition tasks. Sliding window approaches were the earliest action localization approaches that scanned the videos exhaustively to get the video's actions' spatial and temporal coordinates. Some previous action recognition approaches like Silhouette and poses estimation were inspired by object detection frameworks [2]. These frameworks were directly extended to the spatiotemporal scale to localize action. Before Deep Learning approaches came into the picture, handcrafted techniques like Histogram of Oriented Gradient (HOG) [3], Histogram of Optical Flow (HOF) [4], Extended SURF(ESURF) were prevalent [5]. Although these approaches were robust to background noise, change in illumination, and video clutter, they lacked semantics and discriminative capacity.

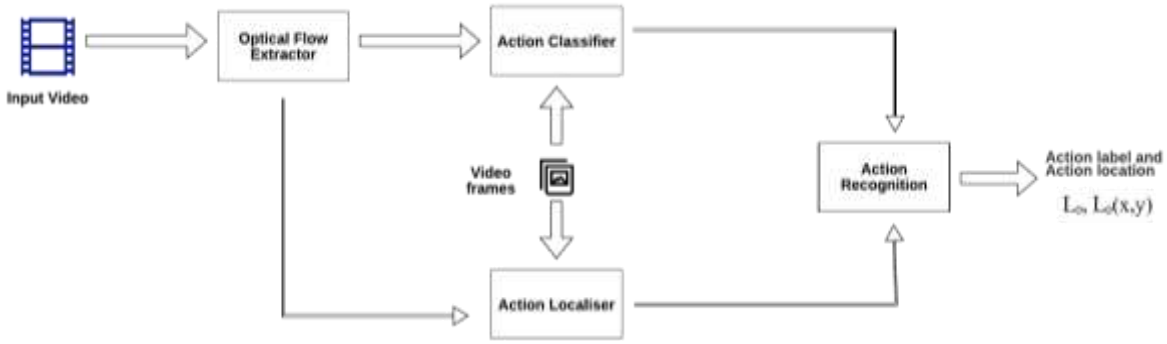


Fig. 1. Action Recognition - Steps Involved.

The purpose of this paper is to analyze the various deep learning architectures for action recognition techniques. It focuses on visual sensor-based methods. The paper has elaborated on action classification, temporal/spatiotemporal action detection, and localization techniques. Section 2 describes the various datasets available for action classification, recognition, detection, and localization. Section 3 explores the different proposed methodologies for action classification tasks. Section 4 delves into the existing approaches for temporal action detection, and Section 5 discusses the methods proposed for spatiotemporal action detection, respectively. Section 6 concludes the whole paper.

II. DATASETS

An estimation says that there are over 1000 human action categories. A variety of studies have been conducted to create datasets that can help us overcome the challenges posed by human action recognition. Action recognition and localization is a widely studied problem. The key challenges associated with this field have been variations in human posture, scaling, pixilation, speed, background clutter, and occlusion. Low-grade and insufficient datasets lead to challenges such as prediction of wrong action class, incorrect spatial or temporal action localization, and inability to detect more than one action in a frame. Table I lists some of the most used datasets for performing action localization and recognition tasks and compares them based on several action classes, data size, nature of video clips, and their aim.

Earlier datasets contained very few action classes. UCF sports has ten action classes: Golf Swing, Lifting, Running, SkateBoarding, Kicking, Diving, Swing-Bench, Swing-Side, Riding Horse, and Walking [6]. UCF sports is introduced, which mainly comprises the video sequences featured on television channels BBC and ESPN.

Various datasets are not realistic, and the action classes are also significantly less. K. Soomro, A. Zamir, and M. Shah [7] targeted these issues and proposed a new dataset, UCF101. It consists of 101 action classes, 13000 vid clips, 27 hours of video clips. Also, the video clips in this dataset are more realistic as they are not recorded in controlled environments, which is essential for training a model which performs well in the real world. However, there is not much variation in the video clips for a particular action class in UCF101.

TABLE I. DATASETS USED FOR ACTION RECOGNITION

Datasets	Number of action classes	Data size	Trimmed/Untrimmed	Year of release	Main Sources
UCF sports	51 action classes	6849 video clips	Trimmed	2008	BBC Motion Gallery and GettyImages
HMDB51	51 action classes	6849 video clips	Trimmed	2011	The Prelinger Archive, YouTube, and Google videos.
UCF101	101 action classes	13320 video clips	Trimmed	2012	YouTube
JHMDB	21 action classes	928 video clips	Trimmed	2013	The Prelinger Archive, YouTube, and Google videos
Thumos15/14	101 action classes	18,420(thumos15), 15,906(thumos14)	Untrimmed	2015(v15), 2014(v14)	YouTube
ActivityNet	200 action classes	9682 video clips(v1.2), 19,994 video clips(v1.3)	Untrimmed	2016(v1.3), 2015(v1.2)	
Kinetics 400	400 action classes	300k video clips	Trimmed (10s)	2017	YouTube
Kinetics 600	600 action classes	500k video clips	Trimmed(10s)	2018	YouTube
Kinetics 700	700 action classes	650k video clips	Trimmed(10s)	2019	YouTube

Some of the datasets focused on increasing the robustness of various action recognition models by exploring under numerous conditions like the movement of the camera, angle, and position of viewpoint, quality of the video, and occlusion. Human Motion Database (HMDB51) [8], dataset focuses on features mentioned above. At least two observers validate the clips of the datasets to establish consistency. The dataset also contains metadata like the number of actors involved, viewpoint, presence or absence of motion of the camera, and category labels.

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. Black [9] proposed JHMDB, "joint-annotated HMDB." This dataset annotates human joints in the HMDB dataset. However, it contains lesser action categories as compared to HMDB51. Their main objective is to understand what features improve the efficiency of action recognition algorithms primarily. They find that high-level pose features are more efficient for capturing actions in videos than low/mid-level features. J-HMDB is beneficial for linking low-to-mid level features with high-level poses. As higher-level pose features need the information of joints. This provides richer information and enables more complex models.

Thumos14 [10] is a dataset used to detect and recognize actions in realistic untrimmed videos with a standard protocol for evaluation. This dataset's action classes are from UCF101, which are mainly divided into five categories - Body-Motion Only, Human-Object Interaction, Human-Human Interaction, Sports, and Playing Musical Instruments. After this, Thumos15 introduces background videos that do not contain the target action with multiple actions in the same video. This further increases the complexity of the dataset.

F. Caba, V. Escorcia, B. Ghanem, and J. Carlos [11] introduced ActivityNet, which has more action categories. Most significantly, ActivityNet has an organized set of activities according to social interactions and where they usually occur. Some of the classes of action in the dataset include - Household, Caring and helping, Personal care, Work-related, Eating and drinking, Socializing and leisure, Sports, and exercises. ActivityNet has the following applications - untrimmed video classification, trimmed activity classification, and activity detection. ActivityNet benchmark has rich semantic taxonomy and aims at covering daily activities performed by humans on an average. Results show that ActivityNet opens new challenges in understanding and recognizing human actions.

Just like various action recognition algorithms are inspired by multiple object detection algorithms. Similarly, some of the datasets are inspired by image datasets. ImageNet inspires kinetics dataset for action classification purposes. The kinetics project aimed to get the same number of action classes as image classes in ImageNet [12]. There are four versions of the kinetics dataset: kinetics 400, kinetics 600, and kinetics 700. Kinetics 400 contains 10 seconds trimmed video clips and a variation in resolution and frame rate having at least 400 clips of each action. Some of the parent action classes in kinetics 400 are arts and crafts, auto maintenance, ball sports, cleaning, dancing, electronics. This dataset can also be used for multi-modal analysis. Kinetics dataset is better than HMDB and UCF datasets due to more action classes and a wide range of actions.

The AVA-kinetics dataset [13] contains 624,430 unique frames and 238,906 unique videos. Some of the selected action classes include swimming, swimming backstroke, swimming breaststroke, swimming butterfly stroke, pushing a wheelchair, giving or receiving awards, punching bag.

III. Deep Learning for Action Recognition

Andrej Karpathy et al. [14] introduced Single Stream Deep Neural networks for action recognition. They proposed and tested four different single stream architectures: Single Frame, Late Fusion, Early Fusion, and Slow Fusion. Single Stream Networks can be induced with information from other models trained on larger datasets to obtain better results. Another significant advantage is that these models do not require the calculator of optical flow as the input includes only RGB images. Therefore, these models can be used for real-time purposes. However, these models were not able to effectively capture the motion features.

To overcome this shortcoming, K. Simonyan and A. Zisserman [15] brought forward the concept of Two-Stream Networks. The Two-Stream Network has two different architectures to individually process the temporal and spatial features. One network takes the single video frames as input, and the other will take the optical flow as input. The output of the two networks is then fused to obtain the class scores. Although this model produces state-of-the-art results in terms of accuracy, it has many drawbacks. As both the networks have to be trained separately, it is not end-to-end trainable. It cannot work with small datasets as transfer learning cannot be applied here. Even though the spatial network can derive features from large image datasets, the temporal model needs to be trained on a video dataset. It is also computationally expensive as the optical flow needs to be calculated before being fed into the temporal network.

Later works made use of LSTMs and 3D convolution networks for action recognition. These networks were not only end-to-end trainable but also worked in real-time. The LSTM architecture was first introduced by Jeffrey Donahue et al. [16]. The authors have taken inspiration from the encoder-decoder architecture and extended it for action recognition. The LSTM based network did not get results as good as the two-stream networks but surpassed the single-stream networks. D. Tran, L. Bourdev, R. Fergus, L. Torresani M. Paluri [17] introduced the concept of 3D convolution networks. This model surpassed the two-stream networks in terms of performance.

The coming sections describe the works that use deep learning techniques for action classification, temporal action detection spatiotemporal action detection.

IV. Action Classification

Action classification is the identification of the type of action in a trimmed or untrimmed video. There has been ongoing research on producing efficient methods of classifying actions in a video clip. L. Wang, Y. Qiao, and X. Tang [18] have put forward a novel video representation

known as Trajectory Pooled Deep Convolutional Descriptor (TDD), which considers the advantages of both deep-learned features as well as handcrafted features. Deep architectures are used to learn discriminative Conv feature maps. Trajectory constrained pooling is conducted to concentrate these convolutional features into effectual descriptors. The accuracy of TDDs is enhanced by using two normalization methods, namely channel normalization, and spatiotemporal normalization, to transform convolutional feature maps. This approach has several advantages. The learning process in TDDs is automatic, and the discriminative capacity is higher when compared to handcrafted features. The plans of action of trajectory-constrained pooling and sampling are introduced by considering the temporal dimension's intrinsic characteristics for aggregating the deep-learned features. The shortcoming of this method is that it is computationally expensive.

Many researchers have made efforts to make the process of action classification less computationally expensive. Although two-stream CNNs are quite efficient and are state-of-the-art when it comes to action recognition, they are computationally costly. One of the main reasons for this is the requirement to calculate the optical flow, which has very high computational needs. The two-stream networks consist of two CNN networks. One is the spatial network that takes as input RGB images, and the other is the temporal network that takes the optical flow as input. This process is not only high on computation but is also time taking. To address this problem, B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang [19] introduced Real-time action recognition with enhanced motion vector CNNs. They have replaced optical flows with motion vectors. Like optical flow, motion vectors describe the motion in a video, but unlike optical flow, they are easily obtained directly in the video decoding process. Hence, they can be used alongside deep convolutional frameworks for action recognition tasks. The authors have proposed a mechanism where the RGB images and motion vectors are obtained from the video decoding process and fed into two-stream CNN. Optical flows are very dense and hence are entirely accurate with fewer noise features. Motion vectors are not very precise and consist of a lot of inaccurate movements and noise. To increase the motion vector CNN's performance, the knowledge learned from an optical flow CNN is transferred into a motion vector CNN. Although optical flow needs to be calculated for this procedure, it is still efficient as this calculation is done only while training and not while testing.

H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould [20] introduced dynamic images for action recognition to further reduce the computational costs. Dynamic images are a novel compact representation of the video, which is based on the rank pooling idea and are acquired through the parameters of a ranking system that encrypts the temporal evolution of the video frames. Since it is an image, CNN models can directly be applied to the video data with fine-tuning allowing end-to-end training for action recognition. This approach is efficient and is not time-consuming as the whole video is summarized to an amount of data equivalent to a single frame.

To further reduce the computation costs while maintaining accuracy, Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann [21] have proposed a Hidden two-stream CNN for action recognition. A two-stream network takes as input RGB images as well as optical flows. The hidden two-stream network is designed to input only the video frames and not the optical flow. This allows a 10x faster performance when compared to the traditional two-stream architecture. This approach uses unsupervised methods to predict the optical flow. The flow field between two adjacent frames is generated using CNN. This indicated flow field and a frame are used to reconstruct the previous frame using backward warping. The idea is that if one frame can rebuild the last frame, then the network has learned the representations of some underlying motions of a video.

While some research works focus on reducing the computational costs of a system, others have attempted to increase the networks' accuracy. W. Byeon, Q. Wang, R. Kumar, and P. Koumoutsakos [22] have proposed a fully context-aware system that produces sharp predictions of high visual quality. The previous prediction models based on CNNs, RNNs, or a combination of both, tend to produce blurry results. Some efforts have attempted to address this issue by separating the foreground from the background, adversarial training, or motion flow learning, but have mainly failed to consider the issue that the model is unaware of the complete information. To solve this shortcoming, the authors have proposed a fully context-aware architecture that captures past information using parallel multidimensional LSTM units.

R. Girdhar and D. Ramanan [23] have also tried to improve action recognition accuracy while ensuring that the network size and computational cost will remain unchanged. They have proposed an Attentional Pooling module that can be used as a replacement for the normal pooling operation in any convolutional network. This model is built over a base Resnet architecture. The proposed Attention layer is plugged into the last layer after generating spatial feature maps, which need to be average pooled.

Another major factor affecting the accurate classification of actions on how much information we can gather from the temporal cues available in the video. Ali Diba et al. [24] have introduced new architecture and transfer learning for video classification. The computer vision community has mainly focused on spatiotemporal approaches where the temporal convolutional kernel depths are fixed. This paper has introduced a new temporal layer that models various kernel depths of temporal convolutions, which are embedded into a proposed 3D CNN. The 3D CNN is extended from the 2D DenseNet by including 3D filters and pooling kernels. Most of the researchers working on 3D convnets tend to train them from scratch. This can prove inefficient as they fail to consider the knowledge gained by the 2D convnets. To overcome this issue, this paper has done an effective transfer of knowledge from 2D convnets to 3D convnets. This not only diminishes the computational cost but also makes the system more accurate.

V. Temporal Action Detection

Temporal action detection is another significant yet testing problem that goes one step beyond action classification. Since recordings in real-world applications are generally long, untrimmed, and contain numerous action instances, this issue

requires perceiving action classifications and recognizing each activity occasion's start time and end time. Temporal action detection can help define the temporal bounds of an action sequence and reduce the computation of action classification tasks. Researchers have tried to solve this problem in various ways. G. Yu and J. Yuan [25] proposed a Fast action proposal for human action search and detection. The action proposal is quite challenging as both the appearance and motion cues have to be considered. This paper is targeted at producing action proposals in unconstrained videos. An action proposal is represented by a temporal series of spatial bounding boxes (spatiotemporal video tube) which can locate a single human action. They have established the action proposal generation as a max set coverage problem, and greedy search is employed to maximize the actionness score. Actionness is a measure that quantifies the likelihood of the presence of an action instance at specified locations. This method can be used before the process of action classification to ensure limited computational costs. The action classification system can now focus only on the action proposals rather than on the whole video. This algorithm works well with moving cameras and can detect actions even in cluttered backgrounds.

Numerous researchers make consistent efforts to facilitate accurate and efficient estimation of actionness. L. Wang, Y. Qiao, X. Tang, and L. Van Goo [26] proposed a hybrid fully convolutional network for actionness estimation. They have introduced a novel convolutional network consisting of an appearance FCN(A-FCN), which takes as input RGB images, and a motion FCN(M-FCN) which takes optical flow fields input. These two networks derive information from static appearance and dynamic motion, respectively. The completely convolutional nature of H-FCN permits it to productively handle recordings with subjective sizes. Each FCN is a discriminative system prepared in a start to finish and pixel-to-pixel way. These estimated actionness maps are then fed into detection frameworks for the action detection process.

Previous temporal action localization strategies depend on applying action classifiers at each time area and different transient scales in a temporally designed sliding window. While most approaches for activity detection find it quite hard to produce high accuracy on large-scale video collections due to their high computational complexity, F. Caba, J. Carlos, and B. Ghanem [27] devised a method to extract temporal segments from untrimmed videos with high recall and good precision at a fast rate. A sparse learning frame is generated for scoring transient frameworks as indicated by the fact that they are prone to contain an action. This proposal is then merged into an activity detection framework to enhance the overall performance.

Many researchers understood the importance of performing temporal action localization in untrimmed videos as recordings in genuine applications are typically unconstrained and contain numerous activity cases in addition to background clutter. To address this issue, Z. Shou, D. Wang, and S. Chang [28] proposed an action localization framework using three-segment-based 3D ConvNets. The framework contains three networks, namely, localization network, classification network, and network. The proposal network is used for identifying action sequences in an untrimmed video. The classification network serves as an initiation for the localization network, which fine-tunes the classification network to localize action temporally.

Single-Stream Temporal Action proposals are another method for obtaining temporal action proposals in long, untrimmed videos [29]. While most methods require the video to be divided into short overlapping clips for temporal action localization, SSTs can process a long video in a single stream. Hence, they are much faster than previous models where temporal action proposals are identified from temporal windows and then independently classified. Applying windows at multiple scales is computationally expensive. Hence, SSTs are less exhaustive and generate action proposals in long videos with just a single video pass through the network.

Single-Stream Temporal Action Detection [30] is another example of a network that incorporates Single-Stream Networks. It draws inspiration from object detection algorithms like YOLO and Faster RCNN. It provided an end-to-end approach of action detection in untrimmed videos, claiming that everything happens in a single pass network. Hence, it is very efficient which can operate at 701 frames/sec. The network was trained for thumos14. This model also outperforms other models in detection performance and fps, just like YOLO.

While most works usually involve building frame-level classifiers and passing the video through them multiple times, S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei [31] have designed a methodology for end-to-end learning of action detection that learns to predict the temporal bounds of actions. An RNN based agent decides which frame to analyze next and when to send forth a prediction. This paper puts forth a single network that takes an untrimmed video for input and gives as output the temporal bounds of any detected actions.

F. Heidarivincheh, M. Mirmehdi, and D. Damen [32] proposed an approach wherein, despite localizing the action, they focused on localizing the moment of completion, where it localizes the completed action along with localizing the action. Hidden Markov Model (HMM) and Long-Short Term Memory (LSTM) are used to assess six kinds of actions - switch, plug, open, pull, pick and drink. The model uses supervised learning. Therefore, the annotations of pre-completion and post-completion frames are already available. They also concluded that fine-tuned CNN features give better results than handcrafted features. An action may often be localized in the video, even if it was an incomplete action. In this approach, by targeting the completion of the action, they successfully overcome this problem.

Some other works also focused on detecting complete actions. An end-to-end trainable network proposed by Yue Zhao et al. [33] Structured Segment Networks focused on untrimmed videos does this by implementing both action classifiers and detecting the complete action. This increases the overall accuracy of the model. Their model also includes detecting high-quality proposal generation termed - Dubbed Actionness Grouping (DAG). The limitation which comes to this model is the existence of a large number of unfinished action snippets in temporal boundaries. To overcome this

issue, the model must understand the various stages of an action. They have introduced structured temporal pyramid pooling that produces a global portrayal of the whole proposal and a broken-down discriminative model to order action classifications together, finding whether a particular action is complete. The model is also computationally efficient because they have used a sparse snippet sampling strategy.

VI. Spatiotemporal Action Detection

Spatiotemporal detection is the process of detecting coordinates of action on a spatial as well as temporal scale. Various algorithms devised for 2D images were directly extended to check their accuracy for 3D actions. One such method is Spatiotemporal Deformable Part Models (SDPM) for Action Detection [34]. This approach explores the generalization of deformable part models from 2D images to 3D spatiotemporal volumes to study their effectiveness for action detection in video. In this paper, a deformable part model is generated for each action (spatiotemporal patterns) and from a collection of examples. The proposed spatiotemporal deformable part model (SDPM) stays true to the structure of the original DPM. This model employs volumetric parts that displace in both time and space, which allows it to perform better for intra-class variation in terms of execution and better performance in clutter.

Another approach that extends a two-dimensional object proposal technique is adopted in spatiotemporal object detection methods [35]. This paper presents spatial, temporal, and spatiotemporal pairwise super voxel features to manage the blending process. Also, they propose another effective super voxel method. Experimental evaluation of the complete model shows that this super voxel approach leads to more precise recommendations than utilizing existing cutting-edge super voxel methods. They have built on the approach of S. Manen, M. Guillaumin, and L. Van Gool [36] that uses a randomized superpixel consolidating methodology to get object proposals.

K. Soomro, H. Igrees, N, and M. Shah [37] proposed early predication and localization of action by taking input at relatively more minor video lengths. Action prediction and online localization accuracies improve over time as the number of frames available increases.

Action localization with tubelets from motion [38] considered super voxels instead of super-pixels to produce spatiotemporal shapes, which directly gives us 2D+t sequences of bounding boxes as tubelets in this paper. Their contributions include investigating the selective search sampling strategy for videos and incorporating motion information in various analysis stages. The singularity of the motion is encoded in a feature vector associated with each super-voxel.

G. Gkioxari and J. Malik [39], inspired by the field of object detection in images, propose an approach where motion and appearance are incorporated in two different ways. In this paper, they select the frames with a higher probability of containing a motion or are more useful for detecting the motion in the video. They select candidate regions and employ CNNs to classify them. The idea of eliminating the regions with lower motion saliency significantly decreases the computation time. The two networks - spatial-CNN and motion CNN operate on static cues and motion cues, respectively.

Other approaches adopted for spatiotemporal action localization include techniques employing dense trajectories. APT: Action localization Proposals from dense Trajectories [40] proposes an efficient generation algorithm to handle many trajectories in a video. The dense trajectories are computed for the video's representation; this paper focuses on re-using them for proposal generation. Therefore, this paper introduces the use of dense trajectories for classification as well.

M. Zolfaghari, G. Oliveira, N. Sedaghat, and T. Brox [41] exploits pose, motion, and appearance for action recognition. To integrate them Markov chain model is utilized, which adds cues successively. This helps in the sequential refinement of action labels.

Action Detection by Implicit Intentional Motion Clustering [42] is based on using spatiotemporal trajectory clustering by leveraging intentional movement properties. The calculated movement clusters are then utilized as action proposals for detection. They find that trajectories from deliberate motion are appreciably densely localized in space and time.

Another group of approaches is based on using two-stream networks for spatiotemporal action detection or localization. Various two-stream networks have been tested successfully for action detection and localization. Two-stream networks consist of a spatial network that models appearance, whose input is RGB frames, and a temporal network that models motion. Optical flow or dense trajectories can be used as input for these networks. Real-Time End-to-End Action Detection with Two-Stream Networks [43] proposes a model that integrates the optical flow computation using Flownet2 and then, applying early fusion for the two streams and training the whole pipeline jointly end-to-end. Experimental results prove that training the pipeline together end-to-end with fine-tuning the optical flow for the objective of action detection improves detection performance appreciably. This model is inspired by YOLOv2.

VII. Conclusion

This paper has presented an expanded overview of various works done in action classification, temporal action detection, and spatiotemporal action detection. Although various on-body sensors are used to understand and label human action recognitions, this paper focuses on visual sensor inputs. Video data is available in abundance and can be effectively utilized for action recognition. The process of action recognition comprises two main tasks, namely, action classification and action localization. The former involves assigning labels to instances of action in a video, and the latter defines the temporal and spatial bounds. Action recognition tasks are challenging due to the lack of complete datasets and high computational cost levels. Significant research has made action recognition a less cumbersome process. A concise summary of multiple datasets employed for action recognition has been presented in the paper. The most used datasets are

compared based on several acting classes, data size, nature of video clips, and their aim. Among the available datasets, the Kinetics 600 dataset has the maximum number of action classes. Although this dataset offers high variation in action types, the videos are trimmed and do not depict real-life scenarios. Contrarily, the ActivityNet dataset offers 200 action classes with untrimmed videos and is a better depiction of real-life activities.

Most of the recent algorithms can localize action in long untrimmed videos with limited computational capacities. The creation of better datasets can significantly improve the performance of these algorithms. The introduction of Single Stream Deep Neural Networks profoundly enhanced the performance of action recognition algorithms. Although this was a considerable breakthrough, these networks had trouble capturing the motion features. It was after this invention that deep learning started to be widely used for action recognition purposes. Later, the introduction of Two Stream Networks made it possible to capture the motion features effectively. Even then, these networks still had a shortcoming of not being end-to-end trainable and fast. LSTMs and 3D convolution networks' proposal made it possible to develop end-to-end trainable, real-time action recognition systems. In the future, the performance of action recognition systems can be significantly increased with the creation of publicly available datasets that contain more action classes with untrimmed videos. Recognizing actions for specific use cases would be much more comfortable with the availability of task-specific datasets. Apparent and standardized documentation of the action recognition methodology would further help make more robust models. Considering a broader set of features and input from multiple sensors while creating models will also significantly improve action recognition systems' performance. The utilization of a range of sensors alongside vision based sensors will drastically improve the performance of deep learning models for action recognition purposes.

REFERENCES

[1] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu and Y. Liu, "Deep learning for sensor-based human activity recognition: overview, challenges and opportunities," arXiv preprint, vol. 37, August 2018.

[2] M. Zolfaghari, G. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," In Proceedings of the IEEE International Conference on Computer Vision, pp. 2904-2913, 2018.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," In IEEE computer society conference on computer vision and pattern recognition, vol. 1, pp. 886-893, 2005.

[4] N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance," In European conference on computer vision, pp. 428-441, 2006.

[5] H. Wangl, M. Ullah, A. Klaser, I. Laptev and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," pp. 124.1-124.11, 2009.

[6] M. Rodriguez, J. Ahmed and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," In IEEE conference on computer vision and pattern recognition, pp. 1-8, 2008.

[7] K. Soomro, A. Zamir and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," In IEEE conference on computer vision and pattern recognition, arXiv preprint, November 2012.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Piggio and T. Serre, "HMDB: a large video database for human motion recognition", In 2011 International Conference on Computer Vision," pp. 2556-2563, 2011.

[9] H. Jhuang, J. Gall, S. Zuffi, C. Schmid and M. Black, "Towards understanding action recognition," In Proceedings of the IEEE international conference on computer vision, pp. 3192-3199, 2013.

[10] Y. Jiang, J. Liu, A. Zamir, G. Toderici, I. Laptev, M. Shah and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes,", 2014.

[11] F. Caba, V. Escorcia, B. Ghanem, and J. Carlos, "Activitynet: A large-scale video benchmark for human activity understanding," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 961-970, 2015.

[12] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, C. Vijayanarasimhan and M. Suleyman, "The kinetics human action video dataset," In Proceedings of the IEEE conference on computer vision and pattern recognition, arXiv preprint, 2017.

[13] C. Gu, C. Sun, D. Ross, C. Vondrick, C. Pantofaru, Y. Li and C. Shmid, "Ava: A video dataset of spatio-temporally localized atomic visual actions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6047-6056, 2018.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725-1732, 2014.

[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Advances in neural information processing systems, vol. 27, pp. 568-576, 2014.

[16] J. Donahue, L. Anne, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625-2634, 2015.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," In Proceedings of the IEEE international conference on computer vision, pp. 4489-4497, 2015.

[18] L. Wang, Y. Qiao and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4305-4314, 2015.

[19] B. Zhang, L. Wang, Z. Wang, Y. Qiao and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2718-2726, 2016.

[20] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi and S. Gould, "Dynamic image networks for action recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3034-3042, 2016.

[21] Y. Zhu, Z. Lan, S. Newsam and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," In Asian Conference on Computer Vision, pp. 363-378, 2018.

[22] W. Byeon, Q. Wang, R. Kumar and P. Koumoutsakos, "Contextvp: Fully context-aware video prediction," In Proceedings of the European Conference on Computer Vision, pp. 753-769, 2018.

[23] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," In Advances in Neural Information Processing Systems, pp. 34-45, 2017.

[24] A. Diba, M. Fayyaz, V. Sharma, A. Karami, M. Arzani, R. Yousefzadeh and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification, arXiv preprint, 2018.

[25] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1302-1311, 2015.

[26] L. Wang, Y. Qiao, X. Tang and L. Van Gool, "Actionness estimation using hybrid fully convolutional networks," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2708-2717, 2016.

[27] F. Caba, J. Carlos and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1914-1923, 2016.

[28] Z. Shou, D. Wang and S. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1049-1058, 2016.

[29] S. Buch, V. Escorcia, C. Shen, B. Ghanem and J. Carlos, "Sst: Single-stream temporal action proposals," In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2911-2920, 2017.

[30] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei and J. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," 2019.

[31] S. Yeung, O. Russakovsky, G. Mori and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2678-2687, 2016.

[32] F. Heidarivincheh, M. Mirmehdi and D. Damen, Detecting the moment of completion: temporal models for localising action completion," arXiv preprint, 2017.

[33] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang and D. Lin, "Temporal action detection with structured segment networks," In Proceedings of the IEEE International Conference on Computer Vision, pp. 2914-2923, 2017.

[34] Y. Tian, R. Sukthankar and M. Shah, "Spatiotemporal deformable part models for action detection," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2642-2649, 2013.

[35] D. Oneata, J. Revaud, J. Verbeek and C. Schmid, "Spatio-temporal object detection proposals," In European conference on computer vision, pp. 737-752, 2014.

[36] S. Manen, M. Guillaumin and L. Van Gool, "Prime object proposals with randomized prim's algorithm," In Proceedings of the IEEE international conference on computer vision, pp. 2536-2543, 2013.

[37] K. Soomro, H. Igrees, N and M. Shah, "Predicting the where and what of actors and actions through online action localization," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2648-2657, 2016.

[38] M. Jain, J. Van Gemert, H. Jegou, P. Bouthemy and C. Snoek, "Action localization with tubelets from motion," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 740-747, 2014.

[39] G. Gkioxari and J. Malik, "Finding action tubes," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 759-768, 2015.

[40] J. Van Germert, M. Jain, E. Garti and C. Snoek, "APT: Action localization proposals from dense trajectories," pp. 2-4, 2015.

[41] M. Zolfaghari, G. Oliveira, N. Sedaghat and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," In Proceedings of the IEEE International Conference on Computer Vision, pp. 2904-2913, 2017.

[42] W. Chen and J. Corso, "Action detection by implicit intentional motion clustering," In Proceedings of the IEEE international conference on computer vision, pp. 3298-3306, 2015.

[43] A. El-Nouby and G. Taylor, "Real-time end-to-end action detection with two-stream networks," arXiv preprint, 2018.

Evaluation of Sentiment Analysis based on AutoML and Traditional Approaches

K.T.Y.Mahima[1]
Informatics Institute of Technology
University of Westminster
Colombo, Sri Lanka

T.N.D.S.Ginige[2]
Universal College Lanka
Colombo, Sri Lanka

Kasun De Zoysa[3]
University of Colombo
School of Computing
Colombo, Sri Lanka

***Abstract*—AutoML or Automated Machine Learning is a set of tools to reduce or eliminate the necessary skills of a data scientist to build machine learning or deep learning models. Those tools are able to automatically discover the machine learning models and pipelines for the given dataset within very low interaction of the user. This concept was derived because developing a machine learning or deep learning model by applying the traditional machine learning methods is time-consuming and sometimes it is challenging for experts as well. Moreover, present AutoML tools are used in most of the areas such as image processing and sentiment analysis. In this research, the authors evaluate the implementation of a sentiment analysis classification model based on AutoML and Traditional approaches. For the evaluation, this research used both deep learning and machine learning approaches. To implement the sentiment analysis models HyperOpt SkLearn, TPot as AutoML libraries and, as the traditional method, Scikit learn libraries were used. Moreover for implementing the deep learning models Keras and Auto-Keras libraries used. In the implementation process, to build two binary classification and two multi-class classification models using the above-mentioned libraries. Thereafter evaluate the findings by each AutoML and Traditional approach. In this research, the authors were able to identify that building a machine learning or a deep learning model manually is better than using an AutoML approach.**

***Keywords*—Automated machine learning; sentiment analysis; deep learning; machine learning**

I. Introduction

Machine learning (ML) is a subset of Artificial Intelligence (AI) and it provides a system to learn automatically. Hence the ML becomes a fast-forwarding application and research development area, there were several new libraries introduced to make the developers' life easier. Moreover, present most industries use machine learning to make their customers', users' lives easier. As an example for this, in 2025 researchers predict that revenue of the AI and machine learning-related enterprise application market would be nearly thirty-one thousand millions of US dollars, Fig. 1 shows the revenue generated and expected from machine learning and AI-related applications [1].

Moreover, there were more than 2000 researchers done research related to the machine learning area [2] . With those statistics, it is clear about the importance of machine learning and the ability of the students and developers to enter the machine learning area. Hence, the machine learning area is an area that is updated day by day. At present, one of the most dominating and focuses gained field in machine learning would be Automated Machine Learning (AutoML).

AutoML plays a new era in machine learning and it is currently an explosive subfield with the combination of machine learning and data science. It provides a set of tools to reduce or eliminate the necessary skills of a developer to implement machine-learning models [3]. This AutoML concept was derived because developing a machine learning model by applying traditional machine learning models needs lots of skills, time-consuming, and it is still challenging for experts as well [4] Moreover it is important to note that the Automated Machine learning method was started in the 1990s for commercial solutions by providing selected classification algorithms via grid search [5].

According to the statistics currently, AutoML is been used by almost all who are involving with data science and machine learning area such as domain experts students, and also governments. People who haven't any knowledge in machine learning, students, or beginner developers are mostly using these AutoML libraries. Fig. 2 gives a clear explanation about how the AutoML usage was distributed with the experience level of the students and employees who are working with machine learning and data science [6].

From these statistics, it can clearly understand there are more than 20% of developers and students who have less experience used AutoML libraries. Because of this, there are several automated machine learning libraries were introduced recently as well. These machine learning libraries were introduced for normal machine learning algorithms and deep learning. Among those TPot, HyperOpt Sklearn, and AutoSckit Learn libraries are very popular. These are the automated versions for the well-known machine learning library Sckit learn [5]. Moreover, for the well-known deep learning library Keras there is an automated library named Auto-Karas [7] Present AutoML is used in various areas in data science and machine learning such as image classification, sentiment analysis, and many more. The main goal of this research work is to evaluate the sentiment analysis based on AutoML and Traditional approaches.

Apart from these libraries, there are several cloud-based AutoML platforms. Google provides their own automated AutoML platform in Google cloud platform named Google AuotML [8]. Moreover, Amazon Web Service (AWS) provides are code-free AutoML platform named AutoGluon [9] Present AutoML is used in various areas in data science and

machine learning such as image classification, sentiment analysis, and many more. The main goal of this research work is to evaluate the sentiment analysis based on AutoML and Traditional approaches.

Sentiment analysis is a method in Natural Language processing. With the rapid growth of social media and the digitalize communication media the sentiment analysis plays a major role in the present. Researchers do various kinds of researches in the sentiment analysis area because of the high availability of the datasets. According to the statistics, there were early 1600 research works done related to sentiment analysis in 2016 [10]. Fig. 3 shows the distribution of publishing research papers related to the sentiment analysis area.

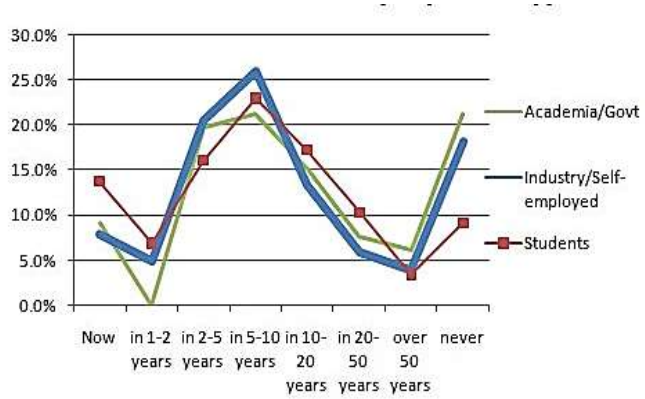


Fig. 1. Increment of the Machine Learning Job Roles in the World.

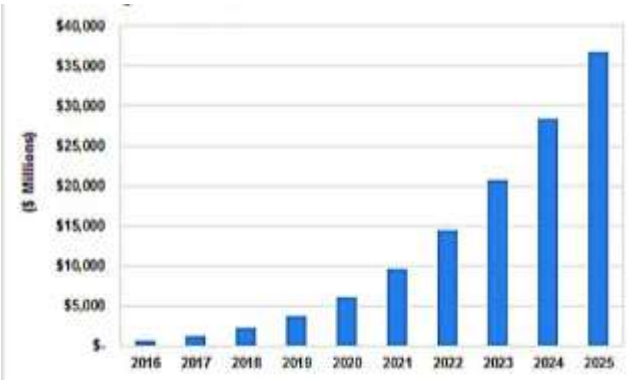


Fig. 2. Distribution of the usage of AutoML with the Experience Level.

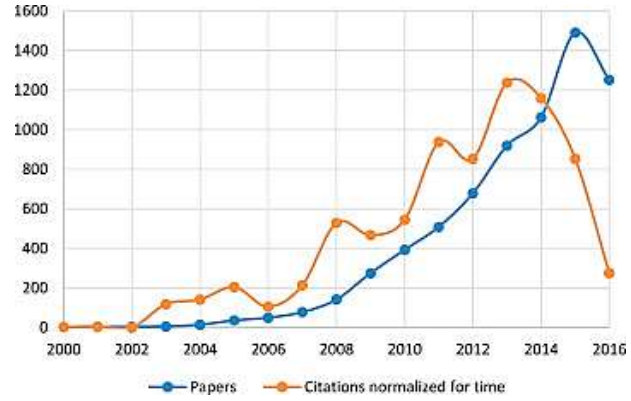


Fig. 3. Distribution of doing Sentiment Analysis Related Research Works.

Because of the rapid growth of the AutoML, this is also used for sentiment analysis projects. This research is evaluated by implementing a sentiment analysis model based on AutoML and Traditional approaches. For that authors chose Python as the main programming language. Here, as the machine learning libraries, authors used Scikit Learn and its automated libraries TPot and HyperOpt SkLearn for the investigation. Moreover, this research evaluated the deep learning approaches as well by using the well-known deep learning library Keras and its automated library Auto_Keras. In the implementation, the authors chose the COVID-19 Tweets dataset, Trip Advisor hotel reviews data set, Spam Message data set, and IMDB Movie Reviews data set which are available in Kaggle to build sentiment analysis based classification models using the above-mentioned libraries. The evaluation of those models was done by comparing the AutoML and Traditional approaches.

Finally from this research. The researchers hope this would be a very useful evaluation for the newcomers to the data science field and students who hope to use AutoML libraries for their projects and this would be a good evaluation to get a proper understanding of the importance of knowing the fundamental knowledge of machine learning. In the next sector, it will discuss several past research works related to the authors' work and how the proposed work differs from those.

II. Literature Review

AutoML is one of the highly focused research areas in Machine Learning. Because of the AutoML, there is considerable growth and interest in doing machine-learning applications. However, the AutoML is a newly introduced evolving technology and lots of researches are being conducted in this particular area hence there could be several pros and cons that could be identified in each AutoML library. Moreover, when using AutoML libraries for each sector such as Image Classification, Time Series based Predictions, or Sentiment analysis, those libraries performance can be varied. Therefore, researchers did several researches to evaluate these AutoML libraries.

A research study named Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools by several researchers from the USA did a great evaluation of AutoML tools used in present. In this study, they investigate AutoML tools like TPot, Auto weka, and several other tools. Moreover, they evaluate AutoML platforms in clouds such as H2O AutoML, Google AutoML. Here they evaluate these libraries by implementing binary-classification, multi-class classification, and regression models for different data sets. As the results, they noted that most AutoML tools obtained reasonable results and performance [11].

	Digits			Fashion		
Benchmark	15 min	30 min	1 h	1.5 h	3 h	6 h
AutoKeras CPU	-	-	-	0.887	0.912	0.912
AutoKeras GPU	-	-	-	0.908	0.921	0.916
AutoKeras GPU with Aug.	-	-	-	0.928	0.933	0.930
H2O	0.984	0.986	0.982	0.902	0.902	0.905
TPOT	0.985	0.985	0.987	0.876	0.879	0.882

Fig. 4. Accuracy of Each Model.

Testing the Robustness of AutoML Systems is research by two researchers from the University of Helsinki. There they evaluate the robustness of three main AutoML libraries for image processing models. For the investigation, they used TPot, H2O, and Auto-Keras libraries. For the implementation, datasets, they used two datasets, which contain digits and fashions. As the result, they were able to get more than 80% accuracy rate from all the AutoML libraries with cleaned data. The table in Fig. 4 summarizes the data accuracy percentages they got in each step of the training [12].

A Journal article named AutoML: Exploration v.s. Exploitation was done to investigate whether AutoML libraries are able to achieve better performance when choosing the most promising classifiers for the given data set. For the implementation, the authors of this article used Auto SkLearn, TPot, and ATM libraries. However, as the results, they mentioned that empirical results across those libraries show that exploiting the most promising classifiers does not achieve a statistically better performance [13].

'AutoML: A survey of the state-of-the-art' is another journal article that discusses the performance of the NAS (Neural Architecture Search). NAS can consider as a subset of AutoML when building deep learning models [14]. In the implementation, they got high perplexity values for the automatically generated models when comparing to the human-made models on PTB (Penn Treebank) data set. Further, explained when the perplexity value is high the model's accuracy is low. Fig. 5 shows those results got by them for each automatically generated and human-made model [15].

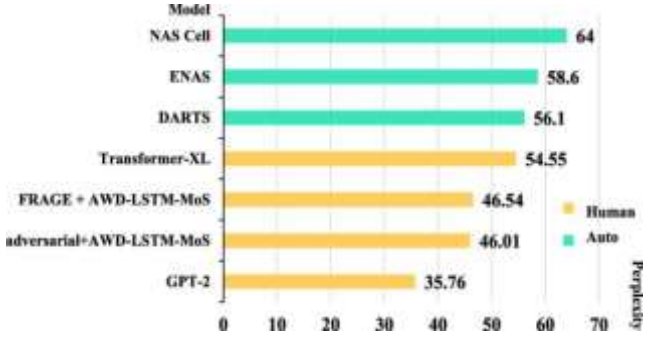


Fig. 5. Perplexity of each Deep Learning Models.

Sentiment Analysis on Google Cloud Platform is a research project, which was done using Google Natural Language API and Google AutoML. The datasets are gathered from Kaggle. As the result, they got nearly 57% accuracy from the google Natural Language API and 90% accuracy from the Google AutoML platform [16]. A web article named Machine Learning-Auto ML vs Traditional methods did an evaluation between python code-based and SAC smart predict system, which is a code-free approach. For the evaluation, the Scikit-Learn library and the Random Forest algorithm were used for the python code-based approach [17].

Another research work named An Open Source AutoML Benchmark introduced an open, ongoing, and extensible benchmark framework that follows best practices and avoids common mistakes. As the open-source AutoML, tools for the benchmark this method used Auto-Weka, Auto SkLearn, TPot, and H2O AutoML libraries. Moreover, for the testing, they used 39 datasets. The main aim of this research work is to compare the accuracy and the performance of each AutoML open-source library. Finally, the findings of the research published their benchmark results as a web-application [18].

So these are some recent competitive works done relevant to the proposed research work. When concentrate on the main goal of this research study, it needs to evaluate the sentiment analysis models based on AutoML and Traditional approaches. Moreover, for the investigation, this research used machine learning automated libraries and deep learning automated libraries. In the next sectors, it will discuss the methodology of the research work, the unique features of it, and the final results gained by the analysis.

III. Methodology

The proposed research work is to evaluate the AutoML and Traditional code-based machine learning on sentiment analysis. For the implementation, the research chose Python as the main programming language since it supports various AutoML libraries. This research work evaluates both machine learning and deep learning AutoML libraries for sentiment analysis. For the evaluation, the researchers chose two main automated machine-learning libraries, which are TPot and HyperOpt Sklearn, and one automated deep learning library, Auto Keras for deep learning. Since these automated machine-learning libraries are based on well-known Scikit learn library the research used Scikit lean and since Auto-Keras is based on Tensorflow Keras deep learning library authors chose Keras to build traditional code-based models.

For the evaluation, the authors implement models for both binary and multi-class classification in sentiment analysis. For that this research used four main datasets. From those, Covid-19 tweets data set [19], Trip advisor hotel reviews [20] data set are used to implement multi-class classification models and Spam message [21], IMDB Movie reviews [22] data sets are used to implement binary classification models. Table I summarizes the datasets and libraries used in this evaluation.

When concentrate on the implementation process as the initial stage this research pre-processed the data by removing null values and special characters. Thereafter authors create word vectors for the texts. When creating word vectors in machine learning models this research used Sckit Learn Tfidfvectorizer method and Python NLTK word_tokenize methods to tokenize the texts and calculate the tf-idf scores in the vector. Moreover, in deep learning models, it used Keras Tokenizer to tokenize words and to make all tokenize sequences into the same length, the Keras pad_sequences method was used.

Thereafter authors build several classification models using Scikit learn and did hyperparameter tunings and record the results got by those. When building the classification models using Scikit Learn Library manually, five main classification algorithms were used. They are XGBClassifier, Random forest classifier, LGBMClassifier, Logistic Regression classifier, DecisionTree Classifier, and. Thereafter using AutoML libraries researchers build Automl models for the same dataset. When building the automated machine learning models authors

allows the library to do further pre-processing. When building the Automl models it trained these Automl models for 100 iterations. For the evaluation of these automated and traditional machine learning models, mainly accuracy and other evaluation techniques like precision, recall, and f1-score were used.

Using the same approach the research builds deep learning models for each dataset using Keras and Auto-Keras. When traditionally building the deep learning models, optimization was manually done by tuning the Hyperparameters. When building these deep learning models mainly Keras LSTM, GlobalMaxPool1D, and Embedding layers were used. In the training process of the Auto-Keras models, also authors trained them for 50 to 100 trials. Finally, for the evaluation of the model, the accuracy and the loss of the model in training and validation times, and other evaluation methods used in classification models were used.

TABLE I. DATASETS AND LIBRARIES

Model Type	Data sets	AutoML Libraries	Traditional Approach Libraries
Multi-Class Classification	1.Covid-19 Tweet Data Set 2.Trip Advisor Hotel Reviews Data set	1.TPot 2.HyperOpt SkLearn 3.Auto-Keras	1.Scikit Learn 2.Keras
Binary Classification	1.Spam Message Data set 2.IMDB Movie reviews data set	1.TPot 2.HyperOpt SkLearn 3.Auto-Keras	1.Scikit Learn 2.Keras

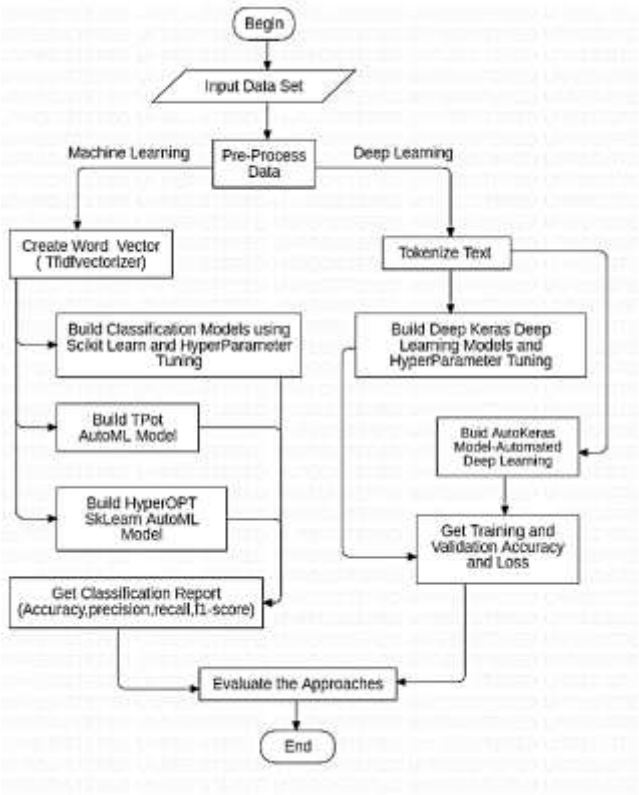


Fig. 6. Flow of the Evaluation for One Data Set.

Fig. 6 describes the flow of the evaluation process for one data set. As a special point when an automated machine learning or deep learning library builds a model, for further evaluation researchers fine-tune that model by manually implementing it using the model parameters output by the AutoML library because in AutoML tools as the accuracy they output the accuracy for overall training, not for the best model.

So this would be the way that the implementation process of the research will continue. This will be a good evaluation to identify the pros and cons of the AutoML vs Traditional approaches of Machine learning. In the next sector, it will discuss the data and results of the research gained in the evaluation process.

IV. DATA AND RESULT

In the implementation, process researchers mainly built two binary classification and two multi-class classification sentiment analysis models using both AutoML and Traditional approaches. As the datasets for the implementation, the research got four main data sets from well know dataset providing site Kaggle. Table II describes each dataset and the number of data available on those. Moreover, in the evaluation process, these datasets were split 70% for training and 30% for the testing.

TABLE II. USED DATASETS AND SIZE

Dataset Name	Size of the Dataset
Covid-19 Tweets Dataset	544735
Trip Advisor Hotel Reviews Dataset	20491
IMDB Movie Reviews Dataset	50 000
Spam Message Dataset	8500

As mentioned in the Implementation sector authors build binary and multimodel classification based on those AutoML and Traditional approaches. When concentrating on the results gained from those as a summary, it can be identified that traditional human-made models perform well and they have high accuracy rates when comparing to the AutoML models in most of the cases. As a special point, the authors identified that when choosing the ML algorithm AutoML libraries are not able to capture the most suitable approach in some cases.

When concentrating on the results got for multi-class classification models using AutoML and Traditional approaches for two datasets used researchers were able to get a 61% accuracy percentage for Trip Advisor Hotel Reviews data set using Logistic Regression Classifier algorithm. Moreover, the deep learning model gets an 82% accuracy percentage. For the AutoML models, it got 75.2% accuracy from the Decision Tree Classification algorithm using TPot and 73.5% from the Random Forest algorithm using HyperOpt Sklearn libraries. Moreover using Auto-Keras as the automated deep-learning library researchers achieved a 74% accuracy rate.

For the Covid-19 Data set which is also used to evaluate the multi-class classification, the research got 71% accuracy from the XGB classifier. And using deep learning researchers got a 72% validation accuracy percentage. Using AutoML libraries it was 60.6% and 60.7% accuracy percentages from the Random

Forest classification algorithm when using HyperOpt-Sklearn library TPot library respectively. When using Auto-Keras as the automated deep learning library authors achieved 57.2 accuracy within 50 trials.

In the implementation, the binary classification in sentiment analysis is also evaluated. When concentrating on the results gained for those datasets, the Spam Message dataset got 97% accuracy from the RandomForest algorithm and 97.6% validation accuracy from deep learning when implemented the models manually. When using the AutoML libraries it got 93.4% accuracy from TPot and 86.7% accuracy from HyperOpt Sklearn library. From Auto-Keras researchers got 89.2% validation accuracy.

Moreover, the IMDB movie rating dataset got 89.7% accuracy from the Logistic Regression classifier and 89.9 validation accuracy from the deep learning model, which create manually. As the results got from automated machine learning from Tpot this research got 76.4% accuracy and 72.1% accuracy from HyperOpt–SkLearn. Here the decision tree classification and random forest classification algorithm were output from those libraries respectively as the highest accurate ones. Moreover, from Auto-Keras, it got 86.5% validation accuracy as an automated deep learning library. Table III summarizes the results got by the authors from all the algorithmic approaches.

According to the results summarized in Table III, it can clearly understand that the performance and the accuracy of the automated machine learning and deep learning libraries are low most of the time when comparing to the ML models made by the authors manually using the Scikit Learn and Keras. Moreover, the accuracy percentages for from TPot and HyperOpt-Sklearn libraries are also quite similar. However, the algorithms suggested by them were different. When building the automated models full control was given to the library for further pre-processing. As an example, the HyperOpt-Sklearn library did Scaling and did dimensional reductions using PCA for the datasets.

TABLE III. RESULTS OF EACH MODEL

Dataset			Trip Advisor Hotel Reviews Dataset	Covid-19 Tweets Dataset	Spam Message Dataset	IMDB Movie Reviews Dataset
Model Type			Multi-Class Classification	Multi-Class Classification	Binary Classification	Binary Classification
Traditional Approach	XGBClassifier	Accuracy	0.57	0.70	0.973	0.82
		Precision	0.50	0.68	0.99	0.83
		Recall	0.43	0.61	0.81	0.83
	Scikit-Learn RandomForest Classifier	Accuracy	0.52	0.65	0.974	0.84
		Precision	0.47	0.76	0.99	0.85
		Recall	0.31	0.51	0.82	0.85
	Scikit-Learn LGBMClassifier	Accuracy	0.59	0.57	0.973	0.86
		Precision	0.54	0.53	0.96	0.87
		Recall	0.49	0.49	0.83	0.87
	Scikit-Learn Logistic Regression	Accuracy	0.61	0.65	0.96	0.89
		Precision	0.56	0.78	0.99	0.90
		Recall	0.50	0.49	0.75	0.90
	Scikit-Learn Decision Tree Classifier	Accuracy	0.49	0.58	0.96	0.72
		Precision	0.40	0.38	0.92	0.74
		Recall	0.34	0.43	0.84	0.72
Automated Machine Learning	TPot	Algorithm	Decision Tree Classification	Random Forest Classification	Random Forest Classification	Decision Tree Classification
		Accuracy	0.752	0.607	0.934	0.764
	HyperOpt-SkLearn	Algorithm	Random Forest Classification Pre-processed using Min-Max Scalar	Random Forest Classification Pre-processed using Min-Max Scalar	Random Forest Classification Pre-Processed using PCA	Random Forest Classification Pre-Processed using Min-Max Scaler
		Accuracy	0.735	0.606	0.867	0.721
Deep Learning	Keras	Accuracy	0.82	0.71	0.976	0.897
		Loss	0.45	0.73	0.368	0.262
Automated Deep Learning	Auto-Keras	Accuracy	0.74	0.53	0.89	0.865
		Loss	0.71	0.92	0.371	0.401

Fig. 7 shows how the HyperPot-Sklearn library scaled the data before training the models.

Moreover, Fig. 8 shows how the HyperOpt Sklearn library uses dimension reduction using principal component analysis (PCA) for the spam message dataset. In addition, for all the four datasets, it can see the HyperOpt-Sklearn library suggests the Random Forest Classification algorithm with a pre-processing technique. However, according to the results, it can clearly understand in some cases these AutoML libraries are not able to capture the highest accurate ML algorithm.

So these are the results got in the evaluation process. In the next sector, it will discuss these results, what researchers could come up with, and proposed about the usage of the automated machine learning and deep learning libraries for sentiment analysis.

```
{'learner': RandomForestClassifier(max_features=0.93237106412596
52, n_estimators=54,
                       n_jobs=1, random_state=2, verbose=False),
'preprocs': (MinMaxScaler(feature_range=(0.0, 1.0)),), 'ex_prepr
ocs': ()}
```

Fig. 7. Pre-Processing in HyperOpt Sklearn.

```
{'learner': RandomForestClassifier(criterion='entropy', max_features=0.69395625
12143973,
                       n_estimators=549, n_jobs=1, random_state=3,
                       verbose=False), 'preprocs': (PCA(n_components=28),), 'ex
_preprocs': ()}
```

Fig. 8. Dimension Reduction Done by HyperOpt Sklearn Library.

V. Discussion

The main purpose of this research is to evaluate the sentiment analysis based on AutoML and Traditional code-based approaches. The implementation of the evaluation was done by using both machine learning and deep learning for multi-class sentiment analysis and binary sentiment analysis. The results gained from those implementations were discussed in the previous sector.

According to the results, researchers can propose several suggestions. The first thing is as students or beginner developers in ML and data science area using a traditional code-based approach is better when comparing to using the AutoML libraries. However, for binary classification models, do not see much risk of using AutoML libraries. Moreover, when concentrate on the results got from Auto-Keras, its accuracy percentages were lower than the traditional approach, and it will perform quite similarly to the traditional approach. Moreover, in the evaluation process, it was identified that the performance of the AutoML libraries are depending on the dataset as well.

Another thing proposed from this evaluation would be, knowing hyperparameter tuning, pre-processing, and knowledge about machine learning algorithms are musts for a beginner when implementing sentiment analysis models. As mentioned earlier in the implementation process authors built five main classification models and did several hyperparameter tunings for those. Moreover, pre-processing data is very useful when building sentiment analysis models. That is one reason why researchers were able to get high accuracy rates when comparing to the accuracy rates got from the AutoML models. In the automated libraries, it was noticed they perform several pre-process techniques such as Min-max Scaler and Principal Component Analysis (PCA).

Moreover, the authors proposed that if someone used the AutoML platform and got a model, he or she must do some tunings for that model and try some similar approaches for that. As an example when the Auto-Keras library returns a model it is good to change, the layers of that model and evaluate it. So these are the main suggestions that are proposed from this evaluation for those who try to use AutoML libraries for sentiment analysis. Moreover, these suggestions and evaluations would be useful in image classification as well.

In the implementation, process of the paper researchers did not face legal or social issues. When making the evaluations to improve the reliability of the findings authors used four data sets evaluated both multi-class classification and binary classification in sentiment analysis. Moreover, the authors evaluate both automated machine learning and deep learning libraries as well to improve the reliability and the validity of the evaluations.

VI. Conclusion

This research was done to evaluate the Sentiment analysis based on AutoML and Traditional code-based approaches. And finally, the research proposed several suggestions for anyone who is going to use AutoML libraries for sentiment analysis. In this research work, the following limitations were identified.

1) Choose only two main automated machine learning libraries, which are very popular.

2) For the investigation, the authors used four datasets.

3) There are code-free automated ml libraries as well. This research ignored those.

When concentrating on the future enhancements of this research author focused on following.

1) Evaluate the cloud-based AutoMLau methods such as Google AutoML.

2) Evaluate other AutoML libraries such as H2O.ai.

3) Evaluate the performance of the AutoML and Traditional approaches in other sectors is machine learning as well. As an example Image classification, Time series analysis can be introduced.

4) Find a method to propose the most suitable algorithmic approach for the given project by the user by analyzing past ML projects.

The authors hope that this research is useful to get a clear underrating of using AutoML in sentiment analysis and the importance of the traditional code-based approach in ML. Moreover, this research shows the importance of knowing the hyper-parameter tuning and other basics of machine learning. According to the results of the research, it is clear that using AutoML platforms is not suitable for every time and it is better to use them to estimate the proper algorithmic approach only. So these are the main findings and values gained in this research.

ACKNOWLEDGMENT

The authors would like to thank all who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. Finally, we would like to thank the University of Colombo, School of Computing for funding this research.

REFERENCES

[1] Intel Corporation (INTC), "Intel FPGAs Powers Microsoft's Real-Time Artificial Intelligence Cloud Platform," 24 08 2017. [Online]. Available: https://seekingalpha.com/article/4101617-intel-fpgas-powers-microsofts-real-time-artificial-intelligence-cloud-platform. [Accessed 21 12 2020].

[2] J. X. Z. Dai Xilei, "A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings," Energy and Buildings, p. 8, 2020.

[3] J. Brownlee, "Automated Machine Learning (AutoML) Libraries for Python," 18 09 2020. [Online]. Available: https://machinelearningmastery.com/automl-libraries-for-python/. [Accessed 22 12 2020].

[4] Eduonix , "A Brief Introduction to AutoML!," Eduonix , 31 12 2019. [Online]. Available: https://blog.eduonix.com/artificial-intelligence/brief-introduction-automl/. [Accessed 22 12 2020].

[5] T. Dinsmore, "Automated Machine Learning: A Short History," 30 03 2016. [Online]. Available: https://www.datarobot.com/blog/automated-machine-learning-short-history/. [Accessed 22 12 2020].

[6] G. Piatetsky, "When Will AutoML replace Data Scientists? Poll Results and Analysis," 03 2020. [Online]. Available: https://www.kdnuggets.com/2020/03/poll-automl-replace-data-scientists-results.html. [Accessed 22 12 2020].

[7] J. ,. S. H. Haifeng, "Auto-Keras: An Efficient Neural Architecture Search System," in 5th ACM SIGKDD International Conference, 2019.

[8] Google, "Cloud AutoML," Google, [Online]. Available: https://cloud.google.com/automl. [Accessed 31 12 2020].

[9] T. A. a. R. B. Abhi Sharma, "Code-free machine learning: AutoML with AutoGluon, Amazon SageMaker, and AWS Lambda," Amazon Web Service, 31 07 2020. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/code-free-machine-learning-automl-with-autogluon-amazon-sagemaker-and-aws-lambda/. [Accessed 31 12 2020].

[10] D. G. M. K. Mika V. Mäntylä, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," Computer Science Review, vol. 27, pp. 16-32, 2018.

[11] J. G. ,. K. H. C. B. B. R. F. Austin Walters, "Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence , Portland, OR, United States, 2019.

[12] J. N. T. M. T Halvari, "Testing the Robustness of AutoML Systems," Electronic Proceedings in Theoretical Computer Science, vol. 319, pp. 103-116, 2020.

[13] H. a. A. E. Eldeeb, "AutoML: Exploration v.s. Exploitation.," ArXiv, vol. abs/1912.10746, 2019.

[14] A. B, "Neural Architecture Search (NAS)—AutoML Process," 12 09 2019. [Online]. Available: https://abinesh-b.medium.com/neural-architecture-search-nas-the-future-of-deep-learning-4b35ca473b9. [Accessed 23 12 2020].

[15] K. Z. X. C. Xin He, "AutoML: A survey of the state-of-the-art," Knowledge-Based Systems, vol. 212, 2021.

[16] M. R. Terrence E. White, "SENTIMENT ANALYSIS ON GOOGLE CLOUD PLATFORM," Issues in Information Systems, vol. 21, no. 2, pp. 221-228, 2020.

[17] S. Bhattacharya, "Machine Learning— Auto ML vs Traditional methods," 8 9 2018. [Online]. Available: https://medium.com/@sidbhat/machine-learning-model-development-8a8dbdd953e3. [Accessed 25 12 2020].

[18] E. L. J. T. ,. P. ,. B. Pieter Gijsbers, "An Open Source AutoML Benchmark," in 6th ICML Workshop on Automated Machine Learning , Long Beach, USA, 2019.

[19] C. Lopez, "COVID-19 Tweets Dataset (over 1 billion)," Kaggle, 11 2020. [Online]. Available: https://www.kaggle.com/lopezbec/covid19-tweets-dataset. [Accessed 2020].

[20] Larxel, "Trip Advisor Hotel Reviews," Kaggle, 10 2020. [Online]. Available: https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews. [Accessed 2020].

[21] UCI Machine Learning, "SMS Spam Collection Dataset," Kaggle, 2016. [Online]. Available: https://www.kaggle.com/uciml/sms-spam-collection-dataset. [Accessed 2020].

[22] A. Anand, "IMDB 50K Movie Reviews," Kaggle, 02 2020. [Online]. Available: https://www.kaggle.com/atulanandjha/imdb-50k-movie-reviews-test-your-bert. [Accessed 2020].

Detecting Hate Speech using Deep Learning Techniques

Chayan Paul[1]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, AP, India

Pronami Bora[2]
Department of Electronics and Communication Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, AP, India

***Abstract*—Social networking sites saw a steep rise in terms of number of users in last few years. As a result of this, the interaction among the users also increased considerably. Along with these posting racial comments based on cast, race, gender, religion, etc. also increased. This propagation of negative messages is collectively known as hate speeches. Often these posts containing negative comments in social networking sites create law and order situations in the society, leading to loss of human life and properties. Detecting hate speech is one of the major challenges faced in recent time. In recent past, there have been a considerable amount of research going on the field of detection of hate speech in the social networking sites. Researchers in the fields of Natural Language Processing and Machine Learning have done considerable amount research in in this area. This paper uses a simple up sampling method to make the data balanced and implements deep learning models like Long Short Term Memory (LSTM) and Bi-directional Long Short Term Memory (Bi-LSTM) for improved accuracy in detecting hate speech in social networking sites. LSTM was found to have better accuracy that Bi-LSTM for the data set considered. LSTM also had better values for precision and F1 score. Bi-LSTM only for higher values for recall.**

***Keywords*—Bi-directional Long Short Term Memory (Bi-LSTM); deep learning; hate speech; Long Short Term Memory (LSTM); text classification**

I. INTRODUCTION

Social Networking Sites (SNS) have provided us with easy ways to connect with various people or organization of our interest. Because of the evolution of various technologies like highspeed internet and handheld devices, these sites have reached to the large number of people in the society. Largest chunk of the users in these networks are young. Researchers have grabbed the large collection of data found in various social networking sites and conducted a considerable amount of research in different areas. Sentiment Analysis is one of the leading areas of research which involves a lot of data from social networks. There are a good number of researches done to find out the sentiment related to a specific product or service using data from social networking sites like Twitter [1] [2] [3] [4]. Apart from sentiment analysis there are other subsets of research done using the data from social networking sites; like detecting users with similar interest in a specific product or service [5] [6]; detection of abusive languages in social media [7] [8]. A good number of research works also have been done to improve the methodologies to analyse the data collected from social networking sites [9] [10].

One thing that these networks make possible now a days is direct interaction with various celebrities. An individual can directly interact with a celebrity and share their views. Similarly, various political parties and business houses utilise these networks for reaching out to their target audience. The problem arises when the users' opinion does not match for an issue. These issues can range from political affiliation to religious belief, opinions related to gender, cast and so on. These mismatch in opinion results in exchange of hate full contents in social networking sites. In fact, hate speech and abusive contents have become a current trend in social media sites and these often results to disturbance in the society. There are reports of riots breaking out in different cities where the main source of the spread of riots are found to be social media posts [11], [12]. Intuitively detection of hate speech in social networks become important.

Hate speech can be characterized as exchange of verbal or nonverbal information among the users with intolerance and aggression [13]. Hate speech can be in different forms, like interaction between users on social network which may contain unparliamentary languages. It could also be abusing a person or a certain group of people for their religious belief, their sexual orientation, their race, their political affiliation [14]. Often these exchange of abusive language lowers the self-esteem of the people and may lead to negative impact in the society [15]. Spread of hate speech has become a global phenomenon.

In this paper endeavors to build a deep learning model for classification of social media contents to either hateful or normal. Twitter was chosen as a platform where detection of hate speech was done. Open source dataset available publicly, was collected to train the models. This paper predominantly builds a Long Short Term Memory and a Bi Directional Long Short term Memory using the dataset.

This section of the paper is followed by a related works section, where the existing works in the related areas are discussed. The next section is methodology, where a discussion is presented on the different methodologies used in this paper. Next to methodology section, results obtained in this paper are discussed. The result section also has introductory discussion on different measures used in this paper for presenting the results. After results section, conclusion section presents the concluding remarks.

II. Related Works

The problem of detecting hate speech has been addressed by various researchers in different ways. In general, the problem can be addressed in different ways. One of the possible ways is to develop a pure Natural Language Processing model, which is generally an unsupervised model. So, the detection becomes comparatively easier as there is no need for a labelled data set. In this approach an NLP model can be designed which categorizes whether a sentence contains hate speech or not [16], [17]. In literature there are fewer works which were carried out totally based on pure NLP based concepts. One of the probable reasons is the models are comparatively slower than the models built using Machine Learning or Deep Learning Models.

The machine learning and deep learning models for detection of hate speech needs labelled data set which is used to train the model. A good number of researches has been carried out in this area where the researchers created their own dataset. The general procedure is to collect the data from a social networking site clean the data and then get them annotated by a team of experts who manually annotate if a text contains hateful message or not. Khan et al., conducted a comprehensive survey of machine learning models used extensively in NLP [18]. Ahmed et al. developed a dataset which consists of English and Bengali mixed texts and annotated the tweets as hate speech or non-hate speech [19]. Sahi et al. developed a supervised learning model to detect hate speech against women in Turkish language. They collected tweets mentioning clothing choices of women and used this data to train the machine learning models [20]. Waseem examined the influence of annotators' knowledge on classification model [21] Waseem et al. provided with a data set of 16,000 tweets and they also investigated which features provides the best performance when it comes to classification of hate speeches [22]. Also, there are a good number of works done where researchers take an open source data and try to develop models which are used to detect the hateful message in social networking sites [23] [24] [25].

The research works in some cases went beyond the binary classification of a message into hate speech and non-hate speech and make it multi class classification. Watanabe et al. conducted a study where they used twitter data to create a model which can classify tweets in three classes i.e., clean, offensive and hateful [26]. Kumar et al. developed a model using taking text messages from Facebook which could classify the messages into three different classes i.e., Aggressive, Covertly Aggressive, and Non-aggressive texts [27].

In this paper we collected a data set from Kaggle which contains tweets from American users. We built a deep learning model to classify the tweets into two categories, hate-speech and neutral.

III. Methodology

In this paper we proposed to classify the tweets using a Long Short Term Memory (LSTM) and a Bi Directional Long Short Term Memory (Bi-LSTM). Both LSTM and Bi-LSTM are versions of neural networks, with persistent memories [28].

A. Long Short-Term Memories (LSTM)

These are special types of neural networks which are designed to work well when one has sequence data set and there exists a long term dependency. These networks can be useful when one needs a network to remember information for a longer period. This feature makes LSTM suitable for processing textual data. Fig. 1 shows a typical architecture of an LSTM. As it can be seen in the diagram, an LSTM is a collection of similar cells, whereas each cell processes the input in a specific approach. Apart from the input from external sources, each cell also receives inputs from its earlier cell in the chain. This arrangement of cells, facilitates LSTM to remember earlier information for a longer time.

B. Bi-Directional Long Short-Term Memories (Bi-LSTM)

Normal form of LSTMs can remember or refer to the information which it has traversed till now. But it does not have any evidence about the information present after the point traversed till the point. This becomes a considerable drawback while dealing with sequence data, especially text. Bi-directional LSTM is another version of LSTM which can remember the information from both directions. In Bi-directional LSTM we basically do backpropagation in two ways. Once from the front and once from the back. This process makes Bi-LSTTM a powerful tool for analysing textual data.

C. Data Pre-Processing

We collected a dataset from Kaggle, an open source platform. The labelled data set contained two classes namely hate speech and non-hate speech. Hate speech is denoted as 1 and non-hate speech is denoted by 0. We removed the special symbols from the texts. Then we converted the texts in lower case. We also used stemming to convert the words into their basic words. We checked the dataset for number of data for hate speech and non-hate speech. We found the data set to be highly imbalanced. Fig. 2 represents the bar diagram for two classes. Table I also represents the number of tweets available in both the classes.

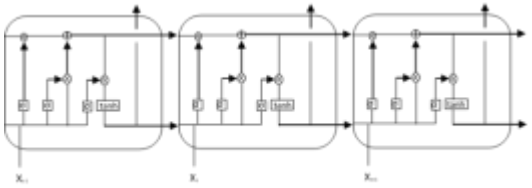


Fig. 1. Architecture of LSTM.

TABLE I. Number of Tweets in Classes

Class name	Number of tweets
Hate-speech (represented by 1)	2242
Non hate-speech (represented by 0)	29720

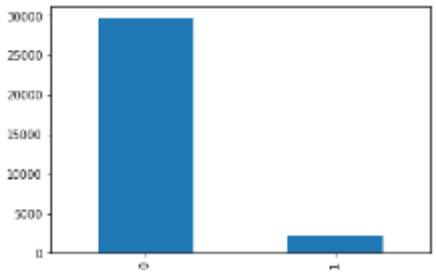


Fig. 2. Bar Diagram Representing Imbalanced Class.

With this state of the data set, if we apply classification algorithms, there is high chance of getting biased results. In this scenario, down sampling can be done to make the majority class equivalent to the minority class. But in this approach, we have risk of losing a large chunk of data which may affect the classification result. Finally, we went for up sampling the minority class, by randomly selecting from the class and adding them back to the data set. This approach provided us with a balanced data set, but the total number of tweets got increased drastically. Fig. 3 represents the balanced data set.

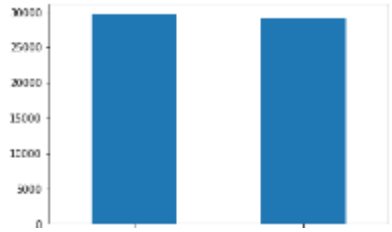


Fig. 3. Bar Diagram Representing Balanced Class.

We divided the data set into training and testing. We kept 67:33 ratio for training and testing. With the training data set we trained an LSTM and a Bi-LSTM. We applied one hot encoding to get the data ready for the algorithms. One hot encoding is a process which converts the text data into numerical data. Each of the words gets a unique numerical representation in one hot encoding. Then we applied padding. Padding is a process which adds zeros to either beginning or ending of sentences for making all the sentences of same length. Then we applied word embedding. Embedding is a process represents each of the words in a higher dimensional space. It is helpful in finding similarity and dissimilarity between the words effectively.

IV. Result

We first computed the confusion matrix for both the models. A confusion matrix presents four different values, namely true positive, true negative, false positive and false negative. True positive means the number of classes which were originally positive, and the model also classified them as positive. True negative means the classes were originally negative and the model also classified them as negative. False positive values are the number of classes which were originally negative, but predicted as positive by the model, and false negative means the classes were originally positive, but predicted negative by the models. Fig. 4 represents the idea of a confusion matrix.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (FP)

Fig. 4. Confusion Matrix.

Fig. 5 and 6 present the confusion matrices for LSTM and Bi-LSTM respectively. In these representations, we presented the values in percentage instead of actual number of classes.

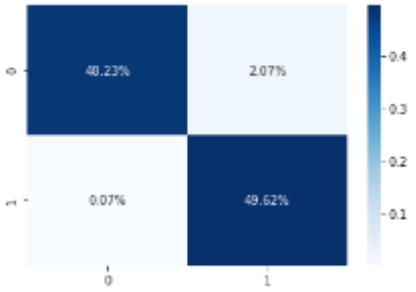


Fig. 5. Confusion Matrix for LSTM.

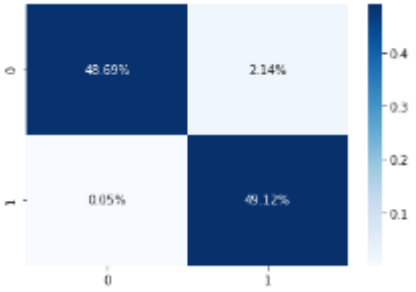


Fig. 6. Confusion Matrix for Bi-LSTM.

From the confusion matrices, we can see there is no considerable difference between the performances of these two models. LSTM has a bit higher false positive in comparison to Bi-LSTM, whereas Bi-LSTM has higher false positive. But it is evident that the differences between the values are very small. We also calculated the other performance measure values accuracy, precision, recall and F1 score. Below we discuss the values in very brief:

A. Accuracy

Accuracy is one of the most widely used performance measures and it is the ratio of total number of entries classified accurately to the total number of observations. For a balanced dataset Accuracy is the measure using which we can compare the performance of an algorithm. In this study, we got a slightly higher accuracy for LSTM, though the difference is very less.

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

B. Precision

Precision is the ratio of entries that are correctly predicted positive to total positive entries. A higher value for precision means low false positive rates. As per the calculations in this study LSTM got slightly higher precision than Bi-LSTM.

$$precission = \frac{TP}{TP + FP}$$

C. Recall

Recall is the ratio of number of positive entries which were predicted correctly to total number of entries in the positive class. It basically reflects the proportion of positive observation which were correctly classified. In this study we can see that Bi-LSTM has better recall in comparison to LSTM.

$$recall = \frac{TP}{TP + FN}$$

D. F1 Score

F1 score is the weighted average of precision and recall, as a result it considers both false negative and false positive. For a problem where the classes are imbalanced, F1 score becomes better performance measure than accuracy. In this study we found the f1 score of LSTM also slightly higher than that of Bi-LSTM.

$$f1\ score = \frac{2 * recall * precission}{recall + precision}$$

We calculated the values for accuracy, precision, recall and F1 score for both the models. The calculated values are presented in Table II.

TABLE II. PERFORMANCE MEASURE SCORES FOR LSTM AND BI-LSTM

Model	Accuracy	Precision	Recall	F1 Score
LSTM	0.9785	0.9598	0.9986	0.9785
Bi-LSTM	0.9781	0.9582	0.9990	0.9781

V. CONCLUSION

The scores calculated for accuracy, precision, and f1 score suggest that LSTM has performed better than Bi-LSTM. But recall score is found to be better for Bi-LSTM than LSTM. Recall basically signifies the ratio of positive classification to total positive classification. Here in this study we considered hate speech as positive class. That means the model has less error in detecting the hate speech. In this context, Bi-LSTM has a slight edge over LSTM. Although, the difference between the scores are really very small to draw any comparison between the two models.

This study can be further extended for real world data set collected from twitter with context to some real events. It will be interesting to see how these models perform on new data set. Attention model is one area which has a good application in NLP, we plan to apply this model in our future works.

REFERENCES

[1] S. Muthukumaran, P. Suresh and J. Amudhavel, "Sentimental analysis on online product reviews using LS-SVM method," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 12, pp. 1342-1352, 2017.

[2] S. A. Devi, P. Sapkota and M. Obulesh, "Sentiment analysis on products using social media," Journal of Advanced Research in Dynamical and Control Systems, pp. 137-141, 2017.

[3] M. Bhargava and D. Rao , "Sentimental analysis on social media data using R programming," International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 80-84, 2018.

[4] C. G. Krishna , D. R. Meka, V. S. Vamsi and K. M. V. S. Ravi , "A survey on twitter sentimental analysis with machine learning techniques," International Journal of Engineering and Technology(UAE), vol. 7, no. 2.32, pp. 462-465, 2018.

[5] P. Jadhav and B. V. Babu, "Detection of Community within Social Networks with Diverse Features of Network Analysis," Journal of Advanced Research in Dynamical and Control Systems , vol. 11, no. 12, pp. 366-371, 2019.

[6] L. P. Maguluri, I. Bhavitha, S. . A. v. Reddy, T. N. Reddy and A. Chowdary , "An efficient method on supervised joint topic modeling approach by analyzing sentiments," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 18, pp. 3219-3230, 2017.

[7] B. R. Rahin, K. K. Prem, N. Danapaquiameq, J. Arumugam and D. Saravanan, , "Blocking Abusive and Analysis of Tweets in Twitter Social Network Using NLP in Real-Time," Bioscience Biotechnology Research Communications, vol. 11, no. 1, pp. 94-103, 2018.

[8] C. Paul, D. Sahoo and P. Bora, "Aggression In Social Media: Detection Using Machine Learning Algorithms," International Journal of Scientific and Technology Research , vol. 9, no. 4, pp. 114-117, 2020.

[9] L. A. Deshpande, and M. R. Narasingarao, "ADDRESSING SOCIAL Popularity in Twitter Data using Drift Detection Technique," Journal of Engineering Science and Technology, vol. 14, no. 2, pp. 922-934, 2019.

[10] S. P. Bhargav, G. N. Reddy, R. R. Chand, K. Pujitha and A. Mathur, "Sentiment Analysis for Hotel Rating using Machine Learning Algorithms," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 6, pp. 1225-1228, 2019.

[11] H. Correspondent, "Facebook played a role in fuelling riots, says Delhi panel," 01 September 2020. [Online]. Available: https://www.hindustantimes.com/cities/facebook-complicit-in-aggravating-n-e-delhi-riots-says-delhi-assembly-panel/story-1HkXrGw4fWSOpoLUrVuCsO.html. [Accessed 03 September 2020].

[12] K. R. Balasubramanyam, "Bengaluru Riots: Karnataka to hold talks with social media giants on filtering fiery contents," 17 August 2020. [Online]. Available: https://economictimes.indiatimes.com/news/politics-and-nation/bengaluru-riots-karnataka-to-hold-talks-with-social-media-giants-on-filtering-fiery-contents/articleshow/77582323.cms. [Accessed 03 September 2020].

[13] K. Sreelakshmi , B. Premjith and K. P. Soman , "Detection of Hate Speech Text in Hindi-English Code-mixed Data," in Procedia Computer Science, Trivandrum, 2020.

[14] J. C. Pereira-Kohatsu, L. Quijano-Sánchez , F. Liberatore and M. Camacho-Collados , "Detecting and Monitoring Hate Speech in Twitter," Sensors (Basel), pp. 1-37, 2019.

[15] A. Gaydhani, V. Doma, S. Kendre and L. Bhagwat, "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach," rXiv preprint arXiv:1809.08651., pp. 1-5, 2018.

[16] G. B. Herwanto, A. M. Ningtyas, K. E. Nugraha and P. T. I Nyoman , "Hate Speech and Abusive Language Classification using fastText," in International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Jetis, Indonesia, 2019.

[17] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the Fifth International workshop on natural language processing for social media., Valencia, Spain, 2017.

[18] W. Khan, A. Daud, J. A. Nasir and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," Kuwait Journal of Science, vol. 43, no. 4, pp. 95-113, 2016.

[19] S. Ahammed, M. Rahman, H. M. Niloy and S. M. H. Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language," in International Conference on System Modeling & Advancement in Research Trends, Moradabad, India, 2019.

[20] H. Sahi, Y. Kilic and R. B. Saglam, "Automated Detection of Hate Speech Towards Women on Twitter," in 2018 International Conference on Computer Science and Engineering (UBMK), Turkey, 2018.

[21] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," in Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, , Austin, 2016.

[22] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proceedings of NAACL-HLT 2016, San Diego, California, 2016.

[23] G. Koushik, K. Rajeswari and S. K. Muthusamy, "Automated Hate Speech Detection on Twitter," in 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019.

[24] T. Davidson, D. Warmsley, M. Macy and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in ICWSM, 2017.

[25] G. K. Pitsilis, H. Ramampiaro and H. Langseth , "Effective hate-speech detection in Twitter data using recurrent neural networks," Applied Intelligence, vol. 48, no. 12, p. 4730–4742, 2018.

[26] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," IEEE Access, vol. 6, pp. 13825 - 13835, 2018.

[27] R. Kumar, A. K. Ojha, S. Malmasi and M. Zampieri, "Benchmarking Aggression Identification in Social Media," in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, New Mexico, USA, 2018.

[28] C. Colah, "Understanding LSTM Networks," August 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 17 09 2020].

Design and Implementation of a Strong and Secure Lightweight Cryptographic Hash Algorithm using Elliptic Curve Concept: SSLHA-160

Bhaskar Prakash Kosta[1]
Research Scholar, Computer Science and Engineering Department, GITAM Institute of Technology, GITAM Deemed to be University, Vishakapatnam, AP, India

Dr. Pasala Sanyasi Naidu[2]
Assoc Professor, Computer Science and Engineering Department, GITAM Institute of Technology, GITAM Deemed to be University, Vishakapatnam, AP, India

***Abstract*—Cryptographic hash function assumes a fundamental job in numerous pieces of cryptographic algorithm and conventions, particularly in authentication, non-repudiation and information trustworthiness administrations. A cryptographic hash work takes a commitment of optional tremendous size message and conveys a fixed little size hash code as a yield. In the proposed work SSLHA-160 (a strong and secure lightweight cryptographic hash algorithm), each 512-digit square of a message is first diminished to 256-bit. A cryptographic hash work takes a contribution of discretionary enormous size message and delivers a fixed little size hash code as a yield. In the proposed work SSLHA-160 (A strong and secure lightweight cryptographic hash algorithm), each 512- digit square of a message is first diminished to 256-bit and afterward partitioned into eight equivalent block of 32 pieces each and each 32-cycle block is additionally separated into two sub-block of 16-piece each. These two sub-blocks go about as two purposes of an elliptic curve, which are utilized for computing another point which is of 16 pieces. The new point esteems are thusly handled to produce message digest. SSLHA-160 is easy to develop, simple to actualize and displays solid torrential slide impact (avalanche), when contrasted with SHA1, RIPEMD160 and MD5.**

Keywords*—Cryptography hash function; message digests; authentication; elliptic curve concepts

I. Introduction

Authentication is a significant idea in information security, and one method of accomplishing this by utilizing hash function. A Hash function is a mathematical function that maps a message of variable size into a small-length esteem called message digest. Message digest is likewise alluded to as a synopsis of the information or unique mark of the information. One direction hash function is a variation of the message authentication code. A single direction hash function, otherwise called a message digest, is a numerical function that takes a variable-length input string and converts it into a small-length arrangement that is computationally hard to alter—that is, going in back direction is impossible create the information from the hash. Hash code is created from all the pieces of the message any blunder in the message can without much of a stretch be distinguished from hash code as little change made in the message brings about huge modification in the hash code. Message confirmation and gadget verification can be accomplished through this hash code or message digest. Likewise message verification is said to secure the respectability of a message, for example the message that is gotten is showing up in a similar frame as send by the sender with no alteration done by clients with vindictive aim. With digital signature well source of messages can be verified. At the point when responsibility for computerized signature mystery key is bound to a particular client, a substantial digital signature gives an affirmation that the message was send by the right client. Sender validation gets significant in monetary and in light weight scenario. This paper is coordinated as follows. Area 2 gives a diagram of cryptographic hash capacities. In Area 3, proposes a strong and secure lightweight cryptographic hash algorithm SSLHA-160, planned dependent on elliptic curve ideas. Area 4 presents results and conversations. The investigation of SSLHA-160 is talked about in Area 5. Area 6 closes by indicating the straightforwardness of calculation.

II. Overview of Cryptographic Hash Function

Hash functions are as of now alluring subject of examination. Especially data security area consistently searches for new ways to deal with plan the safe hash capacities. There are endless hash works that have been created like (Venkateswara Rao Pallipamu, K Thammi Reddy, P Suresh Varma 2014 ASH-160 [1], Venkateswara Rao Pallipamu, K Thammi Reddy, P Suresh Varma 2014 : ASH-512 [2], and Venkateswara Rao Pallipamu, K Thammi Reddy, P Suresh Varma 2016 :ASH-256 [3] Some of the renowned hash work are examined beneath:

A. The MDx Family

The MD calculations are generally used to create an advanced mark from a message. MD2 [14] was created in 1989 for 8-digit encoders. It cushions the message to be encoded until it is a various of 16 bytes long, affixes a 16-byte checksum, and computes the hash. MD4 [12] was created in 1990 for 32-digit encoders. The MD5 [13] message-digest calculation which yield a MD of 128 bit, but it also has shortcoming [19] [20] [22]. Starting at 2019, MD5 keeps on being generally utilized, despite its all-around recorded shortcomings and expostulation by security specialists.

B. The RIPEMD Family

RIPEMD is a gathering of hash work which is created by Hans Dobbertin, Antoon Bosselaers and Bart Preneel in 1992. RIPEMD-160 [4] is a 160-bit cryptographic hash function. It was, intended to be used as a secure replacement for the 128-bit hash function MD4, MD5 and RIPEMD. RIPEMD was developed in the frame work of the EU project RIPE (RACE Integrity Primitives Evaluation, 1988-1992). RIPEMD-160 is a strengthened version of RIPEMD with a 160-bit hash result.

C. The HAVAL Algorithm

HAVAL is a cryptographic hash work. In contrast to MD5, yet like most current cryptographic hash capacities, HAVAL can create hashes of various lengths – 128 pieces, 160 pieces, 192 pieces, 224 pieces, and 256 pieces. HAVAL additionally permits clients to indicate the quantity of rounds (3, 4, or 5) to be utilized to create the hash. HAVAL was broken in 2004. HAVAL was imagined by Yuliang Zheng, Josef Pieprzyk, and Jennifer Seberry in 1992

D. The SHA Family

The Secure Hash Algorithm (SHA) was developed by the National Institute of Standards and Technology (NIST) and published in 1993. A revised version was issued as FIPS PUB 1809-1 in 1995 [5] [6]. This is generally referred as SHA-1. SHA is based on the MD4 algorithm. SHA works by taking care of a message as a bit string of length under 2^{64} pieces, and creating a 160-piece hash esteem known as a message digest. Various version of SHA are SHA-0 (first form, but had loop hole [21], SHA-1 made by National Security Agency but had cryptographic flaw, SHA-2 and SHA-3 [16],[17],[18].

III. Strong and Secure Lightweight Cryptographic Hash Algorithm using Elliptic Curve (SSLHA-160)

The proposed calculation is named as SSLHA-160. This calculation SSLHA-160 (A strong and secure lightweight cryptographic hash algorithm) accepts a message as contribution with a most extreme length of under 2^{64} pieces and creates a 160-piece message digest as yield [10], [11]. First the info message is divided into squares of 512 pieces. This 512 pieces is taken as info, at that point the information is diminished from 512-digit squares to 256-bit blocks. The Hash code creation function acknowledges two sources of info which are 512 pieces square of the message and the initialize MD buffer (fastening variable 160-bits). The cycle comprises of the accompanying advances:

A. Append Cushioning Pieces

The message is cushioned so its length is consistent to 448 modulo 512 (length = 448 mod 512). Cushioning is constantly done, regardless of whether the message is of wanted length. Hence, the quantity of cushioning pieces is in the scope of 1-448. The cushioning comprises of a solitary 1 followed by the important number of 0's.

B. Append Length

A square of 64 pieces which contains the length of the message (prior to cushioning) is affixed to the message. This square is treated as an unsigned 64-bit number (most huge byte first).

C. Initialize MD Cushion

A 160-piece cushion is utilized to hold transitional and end-product of the hash work. The support can be spoken to as five 32-digit registers (S0, S1, S2, S3 and S4) .These registers are instated to the accompanying 32-bit numbers (Hexadecimal qualities):

S0 = 67 45 23 01
S1 = ef cd ab 89
S2 = 98 ba dc fe
S3 = 10 32 54 76
S4 = c3 d2 e1 f 0

These qualities are same as the underlying vector estimations of SHA-1 which are normalized by Federal Information Processing Standards Publications (FIPS PUBS). These qualities are put away in big endian design, which is the main byte of a word in the low-address byte position.

D. Processes Message in 512-Bit Blocks

The Hash code generation method comprises of five sub function. This part is named (H_{SSLHA}-160) in Fig. 1 and its rationale is appeared in Fig. 2. Fig. 1 portrays the general preparing of message to create a message review [7]. The result of the initial two stages (after add cushioning pieces and attach size) outputs a message that is a number multiple of 512-piece long. The extended message is spoken to as the grouping of 512-cycle blocks X0, X1,X2 …X_{L-1}, so the complete size of the extended message is L × 512 pieces (L = the quantity of 512 bit blocks), that is the consequence of different of sixteen 32-bit blocks. Here K speaks to the genuine length of the message in pieces; "IV" is the underlying vector which is utilized to introduce the five 32-cycle registers (S0, S1, S2, S3 and S4). VC1, VC2, … VCq… and VC_{L-1} speak to instate MD(carry vector) which holds middle of the road and eventual outcome of the Hash work, individually. Each round takes two information sources one 512-cycle block (Xq) of the message and a 160 piece convey vector (VCq). Toward the finish of the Lth stage produces 160 piece message digest. At first the given message is isolated into 512-digit blocks, and each square is passed to Hash Code creating function (H_{SSLHA}) as a contribution alongside the 160-piece vector.

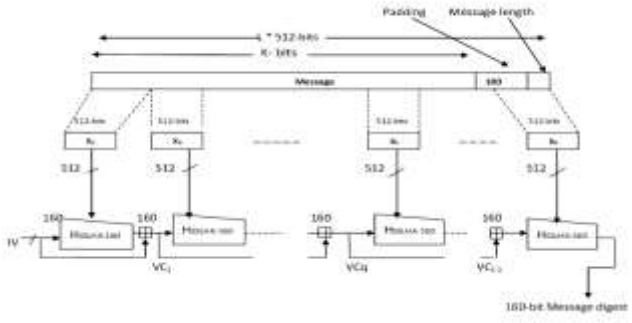


Fig. 1. Hash Code Creation using SSLHA-160.

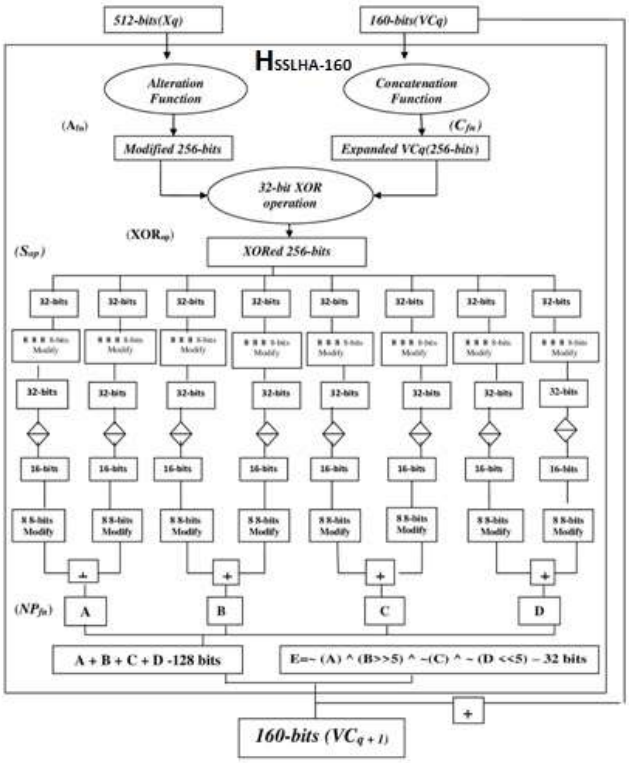


Fig. 2. The Rationale of Hash Work (Compression Work).

The Hash work ($H_{ASSH}160$) rationale is:

Alteration function(A_{fn}) changes over the given 512-bit block into altered 256-digit block.

Concatenation function(C_{fn}) changes over the given 160-piece vector into CONCATENATED 256-bit vector.

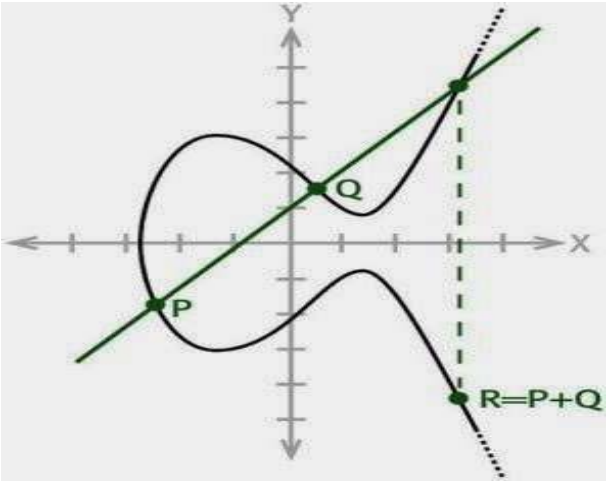
32-digit XOR operation(XOR_{op}) performs XOR procedure on each 32-pieces of changed 256-bit block and concatenated 256-bit vector.

Segregation and Modification operation(Sop) isolates the 256-bit block into 8 sub-block of 32-bits each and then each 32 bit block is separated into four 8-bit block, after that it modifies each eight bit block. This process is repeated for all 32 bit block.

New Point(on elliptic curve) estimation work (NP_{fn}) computes new point on elliptic curve utilizing 32-bit block.

where,

- Xq = the qth 512-cycle square of the message.
- VCq = anchoring variable prepared with the qth square of the message.
- Afn = Alteration function.



P=(xx1,yy1) and Q=(xx2,yy2)

Fig. 3. Elliptic Curve Representation of Two Points.

- Cfn = Concatenation function.
- XORop = XOR(Exclusive-OR) activity performed on each 32-cycle square of the changed 256-digit block (Xq) and relating 32-pieces of the extended 256-bit block (VCq).
- Sop = 256-bit square can be isolated into 8 sub squares of 32-bits each and then each 32 bit block is separated into four 8-bit block, after that it modifies each eight bit block..
- NPfn = first 32-bits further partitioned into 2 sub squares of 16-bits each.

First sub square = 16 pieces, partitioned into two sub squares (8-bits, 8- bits) = (xx1, yy1).

Second sub square = 16 pieces, partitioned into two sub squares (8-bits, 8-bits) = (xx2, yy2).

The qualities (xx1, yy1, xx2, and yy2 as appeared in Fig. 3) are changed over into whole numbers followed by computing a new point on elliptic curve. The above process is redone for remaining seven 32-bit block and the end result is 8 sub block each of 16 bits. Now each of this 8 sub block each of 16 bit is altered ,the first 8-bit of first block is XOR with a 8-bit sub block which are all zeros, the result is stored in first 8-bit sub block. The second 8-bit sub block of first block is added with first 8-bit sub block and result is stored in second 8-bit sub block and the process is repeated for all 8-bit sub block. The result is formatted 128 bit block. Now adjacent sub block each 16 bits are added which results in a sub block of 32 bit and we get four 32 bit sub block. By performing some mathematical operation on above four computed 32 bit sub block and then XOR them results in the fifth 32 bit sub block. This five 32 bit sub block when joined amounts to 160 bit and when added with initialized MD buffer forms the hash code. This 160 bit hash code will be the input to next 512 bit of message i.e. it will act as initialize MD buffer for next 512 bit of the input message. The last 160 bit code generated from last 512 bit of message will be the final hash code which will act as authentication code.

a) Alteration function(A_{fn}): Each 512-bit block of information is separated into 64 sub-blocks involving 8-pieces of each sub-block. One brief exhibit of size 8 (Tempx8[ ]) is taken and instated with zeroes. The modification work comprises of two sub functions as demonstrated as follows:

Sub-function-1: Initially, the above aftereffect of Tempx8[ ] is XOR with the initial 8-pieces of the 512-bit block of message to create the initial 8-pieces of adjusted message. In resulting step this Tempx8[ ] is augmented by 1 and is XOR with the following 8-pieces of the message to create the following 8-pieces of altered message as spoken to in Fig. 4.

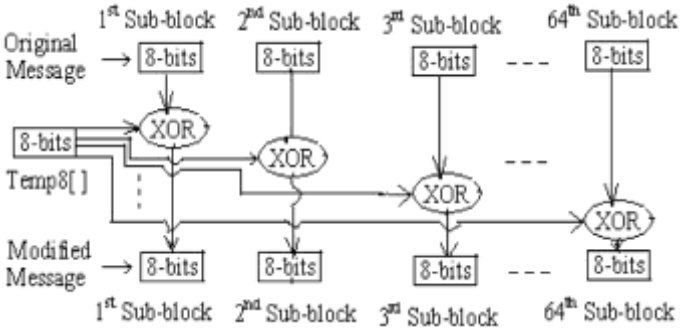


Fig. 4. Operation of Sub-Function-1.

Sub-function-2: The above result is again divided into sub block of 8 bits and 1st 8-bit sub block and middle 8-bit sub block after bitwise complimenting(~) are XOR and the result is again bitwise complimented(~) and stored in a separate array(W2). The above process is repeated for second 8-bit sub block and middle plus one 8-bit sub block and result is stored in second part of separate array(W2). This is repeated for all remaining 8-bit sub block. The result is 512 bit block is reduced to 256 bit block. The modified message as depicted below:

```
int ModInp(vector<int> m, int n)
{
int i,j=0,p,q,k1,T=0;
p=n;
n*=64;
for (i=n;i<n+64; i++)
{
 W[i]=T ^ m[i] ;
 T++;
}
j= ((i+n)/2) ;
q=j;
for ( i=n;i<q+n; i++)
{
 W2[k1]=~(~(W[i]) ^ ~(W[j]));
k1++;
 j++;
  }
 return 0;
 }
```

b) Concatenation function (C_{fn}): The 160-piece Initial Vector (IV) is one of the contributions to the hash code creation function (H_{SSLHA}). It tends to be extended to 256-bits by connecting all underlying vector esteems in roundabout way, which is called Concatenated Vector (VC). The connection cycle is appeared in Fig. 5.

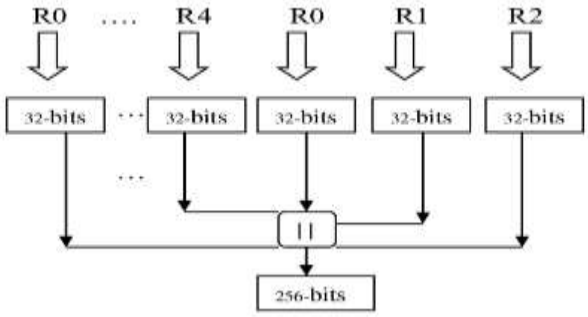


Fig. 5. Concatenation of 160-Bit Block (IV) to 256-Bit Block (CV).

c) 32-bit XOR operation (XORop): XOR activity is performed on initial 32-bit sub-squares of altered 256-cycle block and extended 256-bit block. For this first four 8-bit block are expanded and added so that it becomes 32 bit block. Then this 32-bit block is XOR with first 32-bit of expanded or initialize MD buffer of 256-bit block. This process is repeated for rest of the bits of modified 256-bit and expanded 256-bit block. The result is 256-bit block as shown below:

```
for ( t1 = 0; t1 < 8; t1++)
 {
 W3[t1] = (W2[4 * t1] << 24) + (W2[4 * t1 + 1]
<< 16) + (W2[4 * t1 + 2] << 8) + (W2[4 * t1 + 3]) ;
 if(t1==0)
 W3[t1]=((W3[t1]) ^ (E )) ;
 if(t1==1)
 W3[t1]=((W3[t1]) ^ (D)) ;
 if(t1==2)
 W3[t1]=((W3[t1]) ^ (C )) ;
 if(t1==3)
 W3[t1]=((W3[t1]) ^ (B)) ;
 if(t1==4)
 W3[t1]=((W3[t1]) ^ (A)) ;
 if(t1==5)
 W3[t1]=(W3[t1]) ^ (E) ;
 if(t1==6)
 W3[t1]=((W3[t1]) ^ ( D )) ;
 if(t1==7)
 W3[t1]=(W3[t1]) ^ (C) ;
 }
```

d) Segregation and Modification Operation(Sop): Each 32-digit square of the above outcome is additionally partitioned into 2 sub-squares of 16-bits each. Then each 16-bit block is separated into two 8-bit sub block, each of these eight bits are modified that is First the four bits are taken from the 8-bit value starting from least significant position then it is XOR with a variable1(km) whose value is zero, the result is kept in a variable2(km1) also the value of variable1(km) is updated i.e. it is given the value of variable2(km1). After completing the above step again from 8-bit value(input), four bits starting from most significant bit is stored in variable3(km2) and variable 3 is altered by performing a XOR

with variable1(km). The input i.e. 8-bit value is recomputed by storing the value of variable3(km2)(four bits starting from least significant position) in four bits starting from most significant position and next four bit of initial eight bit(input) stating from least significant position gets the bits from variable2(km1) (four bits starting from least significant position) so this way the 8-bit of input is modified . The same procedure is adopted for remaining three 8-bit values generated from 32 bit that is

B1 = (W3[t] >> 8) & 0xff; →Second eight bit number generated from 32 bit number
A2 = (W3[t] >> 16) & 0xff →Third eight bit number generated from 32 bit number
B2 = (W3[t] >> 24) & 0xff →Fourth eight bit number generated from 32 bit number
A1→ y1

```
int mod_Int2(int y1)
{
 int i,j,km=0,km1=0,km2=0;
 km1=y1 & 0xf;
 km2=(y1 >> 4) & 0xf;
 km1= km1 ^ km;
 km= km1;
 km2= km2 ^ km;
 y1 = (km2<< 4) + (km1);
 return y1;
}
```

A similar system is followed for remaining seven 32-digit blocks.

e) New Point (on elliptic curve) Estimation Work (NPfn): Each 32-bit block is separated into two sub blocks of same length. These sub blocks act as two points of a elliptic curve which are used in new point estimation in elliptic curve [8],[9],[15] as shown in Fig. 3. First by using two points slope of line is calculated the this slope value (M) is used in the calculation of new X and Y axis point on elliptic curve. The resultant new point is 16 bit value i.e X axis is 8 bits and Y axis is 8 bits so both taken together is 16 bit sub block. A total of 8 16-bit sub block are generated after new point calculation. Now this new points are modified by performing XOR on new points. Each 8-bit sub block is XOR with its adjacent 8-bit sub block except first 8-bit sub block which is XOR with a 8 bit sub block of all zeros. After above operation adjacent sub blocks are added in such a way that it becomes 32 bit sub block. Repeating this, results in four 32-bit sub block.

Slope of line in elliptic curve (for Real) is calculated as

$$\text{slope}(\lambda) = \frac{(yy2 - yy1)}{(xx2 - xx1)}$$

where (xx1,yy1) and (xx2,yy2) are points on elliptic curve. New point on elliptic curve for reals are calculated using the following formula

x3(new point) = (slope(λ))2 - xx1 - xx2

and

y3(new point) = slope(λ) * (xx1-xx2) – yy1

the above formula is for the case when x1 != (not equal to) x2 (this is assumed that x1 is not equal to x2)

f) Output: The 160-bit hash code is obtained by attaching four 32 bit sub block obtained in above process and the final 32-bit sub block is obtained by using the above four 32-bit sub block as shown below:

E=~ (A) ^ (B>>5) ^ ~(C) ^ ~ (D <<5)

160-bit message digest = A + B + C + D + E

All hexadecimal numbers are replicated into convey vector (or instate MD cushion), which is the message overview of a given message (if the message size is 512-bit block) in any case the above cycle is rehashed until the last 512-digit square of the message.

Table I representation of input message before padding in hexadecimal and Table II representation of input message after padding in hexadecimal.

TABLE I. REPRESENTATION OF INPUT MESSAGE BEFORE PADDING IN HEXADECIMAL

Ex:. Message "I am from Jabalpur working hard for kits singapur"
Characters count 41.
Bit count 328.

Before cushion (in Hexadecimal representation)

49 20 61 6d 20 66 72 6f 6d 20 6a 61 62 61 6c 70
75 72 20 77 6f 72 6b 69 6e 67 20 68 61 72 64 20
66 6f 72 20 6b 69 74 73 20 73 69 6e 67 61 70 75
72 80 00 00 00 00 00 00 00 00 00 00 00 00 00 00

TABLE II. REPRESENTATION OF INPUT MESSAGE AFTER PADDING IN HEXADECIMAL

After padding:(in Hexadecimal representation)

49 20 61 6d 20 66 72 6f 6d 20 6a 61 62 61 6c 70
75 72 20 77 6f 72 6b 69 6e 67 20 68 61 72 64 20
66 6f 72 20 6b 69 74 73 20 73 69 6e 67 61 70 75
72 80 00 00 00 00 00 00 00 00 00 00 00 00 01 88

g) Alteration function (Afn): Tempx8 = 0 implies a brief exhibit instated with zeroes. XOR activity is performed on every 8-digit square of above advance with Tempx8 by one. First letter(I) decimal value 73 and hexadecimal equivalent is 49. Now 73 is XOR with Temp8 as shown below 73 binary form is 01001001 and temp8 is 00000000 so modified first8-bit block is 01001001 ^ 00000000 = 01001001.

Now Temp8 is incremented by 1 so it becomes 1 and next input decimal value is 32 and hexadecimal equivalent is 20. So the second 8-bit block is modified by XOR it with Temp8 i.e. 00100000 ^ 00000001 = 00100001(decimal equivalent 33) this is the value of second 8-bit sub block and Temp8 is incremented and XOR with third 8-bit sub block value and the process continues till last 8-bit sub block of 512-bit block.

Once the 512-bit block of message is altered it again subjected to modification i.e. the first 8-bit sub block value is complimented and XOR with complimented (mid+1)8-bit sub block value, the result is again complimented and stored in a separate array(W2). Then the second 8-bit sub block value is

complimented and XOR with complimented (middle+2)8-bit sub block value, the result is complimented and stored in the second location of array(W2) and the process continues for rest of 8-bit sub block value till first half of input message(after first step). The result is 512 bit block is reduced to 256 bit block. The modified message as depicted:

The Alteration Step result

W2[0] = ffffff0	W2[1]	=	ffffff90
W2[2] = ffffffcc	W2[3]	=	ffffff92
W2[4] = ffffff94	W2[5]	=	ffffffd0
W2[6] = ffffffd9	W2[7]	=	ffffffc3
W2[8] = ffffff92	W2[9]	=	ffffff8c
W2[10] = ffffffdc	W2[11]	=	ffffffd0
W2[12] = ffffffda	W2[13]	=	ffffffdf
W2[14] = ffffffc3	W2[15]	=	ffffffda
W2[16] = ffffffd8	W2[17]	=	ffffff2d
W2[18] = ffffffff	W2[19]	=	ffffffa8
W2[20] = ffffffb0	W2[21]	=	ffffffad
W2[22] = ffffffb4	W2[23]	=	ffffffb6
W2[24] = ffffffb1	W2[25]	=	ffffffb8
W2[26] = ffffffff	W2[27]	=	ffffffb7
W2[28] = ffffffbe	W2[29]	=	ffffffad
W2[30] = ffffffba	W2[31] = ffffff77		

h) Concatenation function(Cfn): Connecting beginning vector register an incentive in the accompanying design: (link circularly S0 S1 S2 S3 S4 S0 S1 S2)

67 45 23 01
ef cd ab 89
98 ba dc fe
10 32 54 76
c3 d2 e1 f 0
67 45 23 01
ef cd ab 89
98 ba dc fe

i) 32- bit XOR operation (XOR_{op}): XOR activity is performed on each 32-pieces of altered message with relating 32-pieces of extended beginning vector register esteems.

(Modified message) (Expanded vector) (Result)

ef8fcb92	*c*3 *d*2 *e*1 *f* 0 (S4)	2c5d2a62
93cfd8c3	10 32 54 76 (S3)	83fd8cb5
918bdbd0	98 *ba dc fe* (S2)	931072e
d9dec2da	*ef cd ab* 89 (S1)	36136953
d72cfea8	67 45 23 01 (S0)	b069dda9
afacb3b6	*c*3 *d*2 *e*1 *f* 0 (S4)	6c7e5246
b0b7feb7	10 32 54 76 (S3)	a085aac1
bdacb977	98 *ba dc fe* (S2)	25166589

j) Segregation and Modification operation(Sop): Separating the above result (256-bits) into 8 sub-block of 32-bits each and then each 32 bit block is separated into four 8-bit block, after that it modifies each eight bit block as shown in Fig. 6. This process is repeated for all 32 bit block.

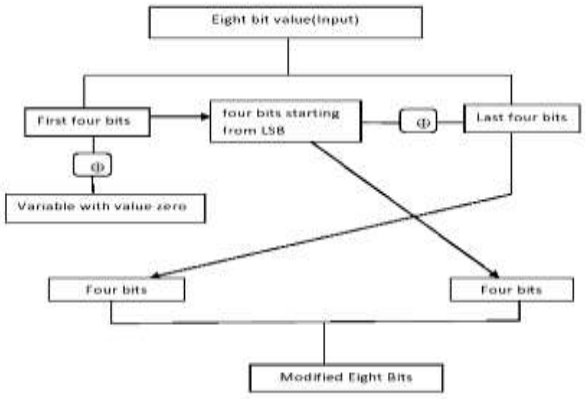


Fig. 6. Alteration of 8-Bit Value by Segregation and Modification Operation.

The Result is

A1[0] is 42	B1[0] is 8a
A2[0] is 8d	B2[0] is ec
A1[1] is e5	B1[1] is 4c
A2[1] is 2d	B2[1] is b3
A1[2] is ce	B1[2] is 77
A2[2] is 21	B2[2] is 99
A1[3] is 63	B1[3] is f9
A2[3] is 23	B2[3] is 56
A1[4] is 39	B1[4] is 0d
A2[4] is f9	B2[4] is b0
A1[5] is 26	B1[5] is 72
A2[5] is 9e	B2[5] is ac
A1[6] is d1	B1[6] is 0a
A2[6] is d5	B2[6] is a0
A1[7] is 19	B1[7] is 35
A2[7] is 76	B2[7] is 75

k) New Point (on elliptic curve) calculation function: Before calculating new point on elliptic curve the 32-bit block is divided into four equal sub-blocks i.e 8-bit block as shown in Fig. 7.

(xx1, yy1) and (xx2, yy2) bits are converted into integers.

```
for ( t = 0; t < 8; t++)
{
A1 = W3[t] & 0xff; // A1=xx1
B1 = (W3[t] >> 8) & 0xff; //B1=yy1
A2 = (W3[t] >> 16) & 0xff; //A2=xx2
B2 = (W3[t] >> 24) & 0xff; //B2=yy2
M=(B2-B1)/(A2-A1); // Slope
```

A3[i]=(M*M)-A1-A2; //A3 = New point and method of calculation.

A3[i]=A3[i]^T1; // New point altered by XOR it with it adjacent (previous) new point except first new point which is XOR with a array of size eight with all zeros.

B3[i]=M * (A1-A2)-B1; // B3= New point and method of calculation.

B3[i]=B3[i] ^ A3[i]; //New point altered by XOR it with it

adjacent (previous) new

point except first new point which is XOR with a array of size eight with all zeros.

```
T1=B3[i];
i++;
}
```

The Result is

```
Calculated New Point value A3[0] is ffffff32
Modified New Point value A3[0] is ffffff32
Calculated New Point value B3[0] is ffffff2b
Modified New Point value B3[0] is 19
Calculated New Point value A3[1] is fffffeee
Modified New Point value A3[1] is fffffef7
Calculated New Point value B3[1] is ffffffb4
Modified New Point value B3[1] is 143
Calculated New Point value A3[2] is ffffff11
Modified New Point value A3[2] is fffffe52
Calculated New Point value B3[2] is ffffff89
Modified New Point value B3[2] is 1db
Calculated New Point value A3[3] is ffffff7e
Modified New Point value A3[3] is fffffea5
Calculated New Point value B3[3] is ffffff87
Modified New Point value B3[3] is 122
Calculated New Point value A3[4] is fffffece
Modified New Point value A3[4] is fffffec
Calculated New Point value B3[4] is fffffff3
Modified New Point value B3[4] is 1f
Calculated New Point value A3[5] is ffffff3c
Modified New Point value A3[5] is ffffff23
Calculated New Point value B3[5] is ffffff8e
Modified New Point value B3[5] is ad
Calculated New Point value A3[6] is 3b3
Modified New Point value A3[6] is 31e
Calculated New Point value B3[6] is ffffff62
Modified New Point value B3[6] is fffffc7c
Calculated New Point value A3[7] is ffffff71
Modified New Point value A3[7] is 30d
Calculated New Point value B3[7] is fffffffcb
Modified New Point value B3[7] is fffffcc6
```

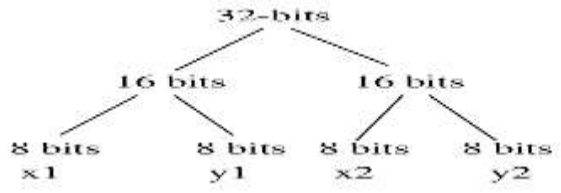


Fig. 7. Portrayal of Four 8-Bit Blocks from 32-Cycle Block.

where in A1 , B1 , A2 , B2 and result A3 , B3 are real numbers. The result A3 and B3 each of 16-bit sub block are altered, then adjacent A3 and B3 are added so that they amount to 32 bit and added with S0(initialize MD buffer) in circular manner to get first 32-bit of hash code of nth 512 bit of message. The process is repeated for other three 32-bit part of hash code. The last 32-bit of hash code is derived by performing some mathematical operation i.e. first 32-bit block A is bitwise complimented(~), the second 32-bit block B is bitwise left shifted by 5 bit, the third 32-bit block C is bitwise complimented(~)and the fourth 32-bit block D is first bitwise right shifted by 5 bit and then bitwise complimented(~) then all A,B,C and D are XORed as shown below.

```
for(i=0;i<16;i=i+2)
 {
 if(i<2)
 AA = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 if(i>=2 && i<4)
 BB = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 if(i>=4 && i<6)
 CC = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 if(i>=6 && i<8)
 DD = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 }
```

```
EE=~(AA) ^ (BB>>5) ^ ~(CC) ^ ~(DD <<5);

printf("\n %08x %08x %08x %08x %08x\n\n", AA, BB,
CC, DD, EE);
```

Result of above operation:

```
AA= 3217f843
BB= 53d9a622
CC= ec1e23ad
DD= 1a7f09c6
EE= 6c89d1e0
```

l) Output: The 160-bit hash code is made by joining four 32 bit sub block obtained in above process and the final 32-bit sub block is obtained by using the above four 32-bit sub block generated in above process as shown below:

```
EE=~(AA) ^ (BB>>5) ^ ~(CC) ^ ~(DD <<5);
```

The value of AA,BB,CC,DD,EE is 3217f843 53d9a622 ec1e23ad 1a7f09c6 6c89d1e0

Then the final hash code for 512 bit block of the message is generated by adding (S0 , S1 , S2 , S3 , S4) to individual 32 bit of above result as shown below:

```
AA = S0 = S0 + AA;
BB = S1 = S1 + BB;
CC = S2 = S2 + CC;
DD = S3 = S3 + DD;
EE = S4 = S4 + EE;
```

final hash code: d3cef4e1 0a4cb54f 84d900ab 1a7f09c5 97a1d6d1

All hexadecimal numbers are replicated into instate MD buffer, which is the message condensation of a given message in the event that it is a 512-digit block in any case the above cycle is rehashed until the last 512-bit block.

For ex: If Elliptic curve new point is (8, 12), then the binary form is

and its modified form is shown below
8 in its binary form is 00001000
12 in its binary form is 00001100.
8 cycle XOR operation is as per the following:

Step 1: Let AA1 be a array of size 8 and duplicate these qualities into AA1.

AA1 [ ] = { 0, 0, 0, 0; 1, 0, 0, 0 };

Step 2: Let Tempx8 be an array of size 8 and introduce with zeroes.

Tempx8 [ ] = {0, 0, 0, 0, 0, 0, 0, 0};

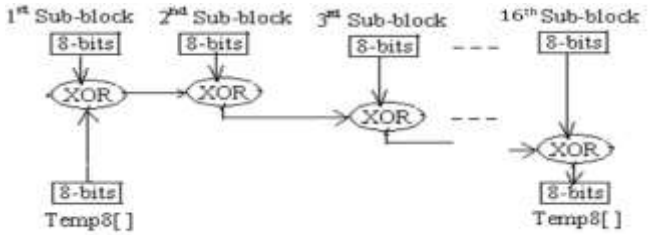


Fig. 8. Operation of Step3.

Step 3: Perform XOR operation on every 8-bits of AA1 and Tempx8:

{0 0 0 0 0 0 0 0}^{0 0 0 0 1 0 0 0} = {0 0 0 0 1 0 0 0}

Now 1st 8-bit sub block is modified to {0 0 0 0 1 0 0 0} and Tempx8[ ]={0 0 0 0 1 0 0 0} and AA1[ ] takes the other value of new point so AA1[ ] = {0 0 0 0 1 1 0 0}. Now XOR operation is done between AA1[ ] and Temp which is shown below.

{0 0 0 0 1 0 0 0} ^ {0 0 0 0 1 1 0 0} = {0 0 0 0 0 1 0 0}

Again the 2nd 8-bit sub block is modified i.e. it becomes {0 0 0 0 0 1 0 0} and tempx8[ ] changes its value i.e. Tempx8 [ ] = {0, 0, 0, 0, 0, 1, 0, 0}; as shown in Fig. 8 and the process continues for all the other seven new point.

Step 4: Above operation results in 128 bits i.e. 8 sub block of 16 bits. Here two adjacent sub block are taken and added in such a way that the result is 32 bit sub block. the process is shown.

```
for(i=0; i<16; i=i+2)
 {
 if(i<2)
 AA = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 if(i>=2 && i<4)
 BB = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 if(i>=4 && i<6)
 CC = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 if(i>=6 && i<8)
 DD = (A3[i] << 24) + (B3[i] << 16) + (A3[i+1] << 8) +
(B3[i+1]);
 }
```

The above process results in four 32-bit sub block. By using this four 32-bit sub block fifth 32 bit sub block is computed which is shown below:

EE=~(AA) ^ (BB>>5) ^ ~(CC) ^ ~(DD <<5)

The result for the message “I am from Jabalpur working hard for kits singapur” is 3217f843 (AA) 53d9a622 (BB) ec1e23ad (CC) 1a7f09c6 (DD) 6c89d1e0 (EE)

After computing AA , BB , CC , DD , EE , this values are added with individual 32 bits of initialize MD buffer (S0 , S1 , S2 , S3 , S4) to generate the new expended value for next 512 bit of the given message. The process is shown below

AA = S0 = S0 + AA;
BB = S1 = S1 + BB;
CC = S2 = S2 + CC;
DD = S3 = S3 + DD;
EE = S4 = S4 + EE;

AA, BB , CC , DD , EE values generated for last 512 bit of the given message will be the hash code for authentication.

The hash code for the message “I am from Jabalpur working hard for kits singapur” is d3cef4e1 0a4cb54f 84d900ab 1a7f09c5 97a1d6d1

IV. Results and Discussion

SSLHA-160, RIPEMD 160, MD5 and SHA-1 when implemented in and run on windows 7 32bit, processor 2.00 Giga Hz with 2 GB of internal memory.

Input String: “The student tried hard but failed in the exam”.

Output (Hash Code): 293967e0 a2141c90 93f12216 b2467106 ed0c49d0.

When the given message is slightly changed, this results in huge alteration in the yield due to the avalanche effect, which is a property of Hash function. For example, when the word exam is changed to rxam i.e. a single letter is changed this produces a hash code which differs from the original hash code by 92 bit out of 160.

Input String: “The student tried hard but failed in the rxam”.

Output (Hash Code): fab3ee1b facc7548 dc39da9e 0afec9be be86d00b

293967e0 :0010 1001 0011 1001 0110 0111 1110 0000
fab3ee1b : 1111 1010 1011 0011 1110 1110 0001 1011
---
3 2 1 2 1 2 4 3
---
a2141c90 :1010 0010 0001 0100 0001 1100 1001 0000
facc7548 : 1111 1010 1100 1100 0111 0101 0100 1000
---
2 1 3 1 2 2 3 1

93f12216 : 1001 0011 1111 0001 0010 0010 0001 0110
dc39da9e: 1101 1100 0011 1001 1101 1010 1001 1110
---
1 4 2 1 4 1 1 1
---
b2467106: 1011 0010 0100 0110 0111 0001 0000 0110
0afec9be: 0000 1010 1111 1110 1100 1001 1101 1110
---
3 2 3 1 3 1 3 1
---
ed0c49d0: 1110 1101 0000 1100 0100 1001 1101 0000
be86d00b: 1011 1110 1000 0110 1101 0000 0000 1011
---
2 2 1 2 2 2 3 3

Total number of bit changed is: 18+14+15+17+17 =81 bit out of 160 bits. When the same experiment was done on SHA-1 it also showed avalanche effect, but only 76 bits changed, in case of MD5 74-bits (57-bits for 128 bits message digest) and in case of RIPEMD-160 82 bits changed as shown in Fig. 9.

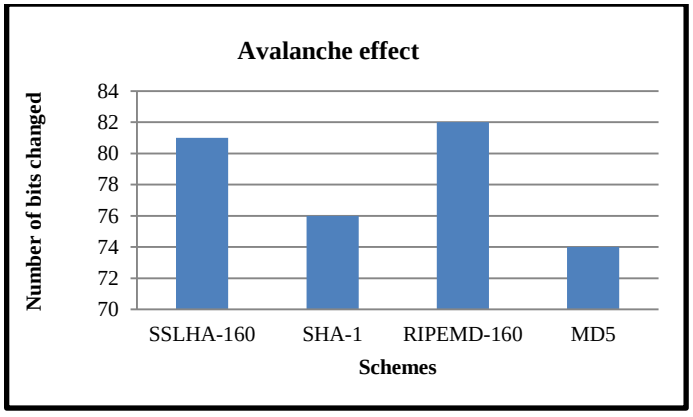


Fig. 9. Avalanche Effect is Demonstration for the Message “The Student Tried Hard but Failed in the Exam” and “The Student Tried Hard but Failed in the Rxam”.

Table III shows the message digest produced by four algorithm (SSLHA160, SHA-1, RIPEMD160 amd MD5) for the message “The student tried hard but failed in the exam”. The aftereffect of Table III is utilized for computing avalanche impact appeared in above Fig. 8. Table IV shows the adjustment in message digest when the message in Table III is modified for example exam is supplanted with dxam, SSLHA-160 produces a message digest with change of all hexadecimal qualities aside from 9,6 and c(starting from left shown in intense), SHA-1 yield is a hash code with change of all hexadecimal qualities aside from 5,7,e,f and 5(starting from left shown in intense) and RIPEMD 160 produces a message digest with change of all hexadecimal qualities aside from 8 and 5(starting from left shown in strong) and Table V shows the adjustment in hash code when the message in Table III is modified for example exam is supplanted with fxam, SSLHA-160 produces a message digest with change of all hexadecimal qualities aside from 2 and e(starting from left appeared in striking), SHA-1 yield is a hash code with change of all hexadecimal qualities aside from 3(starting from leftshown in intense) and RIPEMD 160 produces a message digest with change of all hexadecimal qualities aside from b,0,0 and 5(starting from left shown in strong). So on an average SSLHA-160 shows similar result when any character is changed. The result is better than SHA160 and RIPEMD160 as shown in Fig. 10.

TABLE III. RESULT

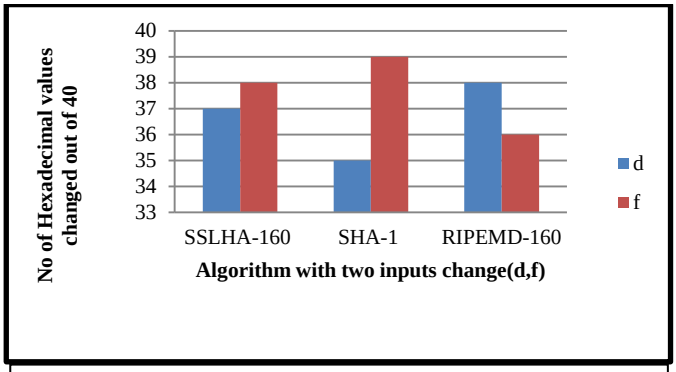
Algo/Input : “The student tried hard but failed in the exam”
SSLHA160: 293967e0 a2141c90 93f12216 b2467106 ed0c49d0
SHA1: d6574b20 5ecbfcec 90161a3f b5fb335a 2d20ed77
RIPEMD160 : 996c99e3 8f901582 3b3bb008 c0f452bf 42ff27bc
MD5 :6fa7ff78 11a4814a d4a0699e 7df98253

TABLE IV. RESULT (IN INPUT EXAM CHANGED TO DXAM)

Algo/Input: “The student tried hard but failed in the dxam”
SSLHA160: d85**9**6efa 5660515c c8a56eca 6692a5d2 9c2**c**50ea
SHA : 2f**5**798c9 f**e**6025a8 a648850e ea**f**427**5**0 474b9bce
RIPEMD160: 51a8d28a **8**6b7344f b54efa6f a1dc**5**609 8f2216fd

TABLE V. RESULT (IN INPUT EXAM CHANGED TO FXAM)

Algo/Input : “The student tried hard but failed in the fxam”
SSLHA160: **2**0a60123 f3c16e3d e144d5a9 03f3c2b3 **e**478e313
SHA1: 8086fa8b 3f5cd5da db0cd33d 5496a8b1 e954ac0b
RIPEMD160: b6363eac 49652c09 9c0**b**dd**0**7 6**0**a0**5**bd6 e3d5759b



Message: "The student tried hard but failed in the exam "

Compared with

d : "*The student tried hard but failed in the dxam* "

f : "*The student tried hard but failed in the fxam* "

Fig. 10. Comparing Three Algorithm with Two Inputs (Exam is Replaced by Dxam and Fxam).

V. Analysis of SSLHA-160

A. Security Analysis

The hash work SSLHA-160 means to satisfy essentially two targets, one is solid avalanche impact and another is one way property (for example irreversible property of hash work). The difference in a digit will mirror the difference in the hash code by performing XOR activities as referenced in Alteration capacity and XOR activity. In any case, XOR is straight, and doesn't forestall differential assaults. Yet, change of single piece mirrors a great deal of progress in hash code (for our example 81 bit is changed). The one way property or irreversibility property of hash value is achieved by new point (on elliptic curve) estimation work as referenced in Section V. The One way property is clarified underneath:

The yield of Section IV Segregation and modification activity (Sop) Each 32-cycle square of the above outcome is additionally partitioned into 2 sub-squares of 16-bits each. Each 16-bits are modified i.e. it is separated into 8-bit each then this 8-bit is altered, then this 8-bits goes about as a point in the elliptic curve, for example initial 8-bits go about as X-pivot worth and second 8-bits go about as Y-hub esteem. The with the this four 8-bit values a new point on elliptic curve is calculated , this new point is of 16-bits again it is partitioned into two 8-bits sub block and each 8-bit sub block value is modified. A similar system is followed for every one of the 32-bit blocks.

New point: The Calculated A3[0] value is ffffff32 and in binary is 00110010

The Modified A3[0] value is ffffff32 and in

binary is 00110010

The Calculated B3[0] value is ffffff2b and in binary is 00101011

The Modified B3[0] value is 19 and in binary is 00011001.

Above way is adopted for calculating new point of remaining sub blocks and then modify it, after modifying the adjacent new point are added such that they becomes 32-bit sub block. Repeating the process for rest of modified new point's results in four 32-bit numbers. By using this four 32-bit sub block i.e. performing some mathematical operation and XOR them results in fifth 32-bit sub block. The outcome is linked and changed over to hexadecimal structure. In this way, it shows the single direction(one way) property of SSLHA-160 hash calculation.

Assaults not related with the calculation are:

Irregular assault- The likelihood of breaking this calculation is 1/2160, the quantity of trails and the expected values are the vital boundaries of this assault.

Birthday assault- The idea driving birthday assault came from a well-known issue from probability theory, which is called birthday paradox. By utilizing this idea assault on hash capacity can be outlined. On the off chance that the length of the Message digest is 160 pieces, at that point there are 2160 prospects. The Cryptanalyst produce two examples which are P1 and P2 from digest, the estimated likelihood of two examples is as per the following:

$$\rho \approx 1 - \frac{1}{e^{\frac{p_1 p_2}{2^{160}}}}$$

Assaults related on the calculation are:

Meet-in-the-middle assault- is a variant of the birthday assault, here aggressor endeavors to locate any two q1 and q2 with the end goal that their hash esteems are equivalent to y = h (q1) = h (q2). This assault is identified with discovering two people with a similar birthday. Let x be the probability that two person birthday are equal and x! be the probability that the birthdays are not equal. Then the probability is defined as

$$1 - x! = x$$

Now suppose that there is only one person, then the probability that his birthday is not same with any one is 1. If there are two person then the probability that they have birthdays on different dates are 1 * 364/365(considering year to be of 365 days). Similarly if three persons are there then the probability that they have birthday on different dates is 1 * 364/365 * 363/365. If we calculate for 9 persons i.e. the probability that nine person have birthday on different dates are:

1 * 364/365 * 363/365 * 362/365 ------9Person

When multiplied we get around 0.9 i.e. 90% which means that the probability that out of nine person two persons having birthday on same date is 10%. Going in the same way when we calculate probability for 23 person having birthday on different dates we get around 49.9% and the probability that two persons out of 23 having birthday on same date will be greater than 50%. One expects lower probability as there are 23*22/2 = 253 pairs of persons.

The birthday paradox can be utilized for assaulting hash capacities. An enemy produces q1 varieties of a sham message and q2 varieties of a veritable message; n is the quantity of

pieces of a hash esteem. At that point, the likelihood of finding a sham message and a certified message as follows:

$$P \approx 1 - (1 \div \left(e^{\frac{q_1 q_2}{2^n}}\right))$$

Remedying block assault - The analyzer takes a message and its hash code on which he attempts to change the block a few times and notices the summary(digest) remaining parts same or not.

Differential assault - The rule of an assault is the investigation of social contrasts among information and yield. An impact happens if the difference is zero.

B. Performance Analysis

These three calculations SSLHA-160, SHA1 and MD5 were tried for comparison dependent on the execution time necessities as shown in Fig. 11, Fig. 12 and Fig. 13. All the calculations have been actualized in C/C++ and run on windows 7 32-bit, CPU 2.00 GHz with 2 GB of internal memory. With the aftereffects of the analysis, it was discovered that SHA1 and MD5 requests more execution time than SSLHA160 to create hash code.

The result in above Fig. 14 demonstrates time taken by three algorithm (SSLHA160, SHA1, and MD5) for generating message digest for the message "I am from Jabalpur working hard for kits singapur".

Fig. 11. Time Taken by SSLHA160 for Executing "I am from Jabalpur Working Hard for Kits Singapur".

Fig. 12. Time Taken by MD5 for Executing "I am from Jabalpur Working Hard for Kits Singapur".

Fig. 13. Time Taken by SHA-1 for Executing "I am from Jabalpur Working Hard for Kits Singapur".

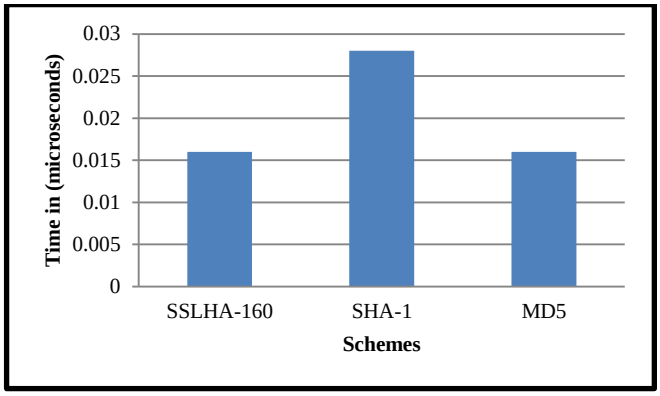


Fig. 14. Examination of SSLHA160, SHA1 and MD5 Concerning Time Taken to Execute the Message "I am from Jabalpur Working Hard for Kits Singapur" i.e. 41 Bytes.

VI. Conclusion

The necessity for new hash plans is expanding to make security viewpoints solid, for example, validation, message honesty and secrecy as for present status of web as web conditions are much of the time evolving. The specialists in cryptography should invest solid energy to concoct better plan measures utilizing which long haul and powerful security can be given by hash capacities. The proposed hash work, A strong and secure lightweight cryptographic hash algorithm (SSLHA-160) is created utilizing elliptic curve ideas which is contrasted with different cryptographic hash capacities. The center qualities of SSLHA-160 are XOR activities, bitwise compliment , bitwise left and right shift and new point estimation(for elliptic curve), which bring about solid nonlinear avalanche impact, expanded dispersion in yield and make differential assaults troublesome. Consequently it is safer and simple.

References

[1] Venkateswara Rao Pallipamu, K Thammi Reddy, Suresh Varma "ASH-160: A novel algorithm for secure hashing using geometric concepts" 2014 Elsevier Ltd.

[2] Venkateswara Rao Pallipamu, K Thammi Reddy, P Suresh Varma "ASH-512: Design and implementation of cryptographic hash algorithm using co-ordinate geometry concepts" 2014 Elsevier Ltd.

[3] Venkateswara Rao Pallipamu, K Thammi Reddy, P Suresh Varma "Design and implementation of geometric based cryptographic hash algorithm: ASH-256" 2016 IJESRT

[4] Dobbertin H, Bosselaers A, Preneel B. RIPMEMD-160: a strengthened version of RIPMMD. In: Gollmann D, editor. Fast software encrzptions, LNCS 1039. Springer-Verlag; 1996. pp. 71 - 82.

[5] Federal Information Processing Standards Publication. Secure Hash Standard (SHS). FIPS PUB; March 2012. pp. 180 - 4.

[6] Federal Information Processing Standards Publication. Secure Hash Standard (SHS).MD 20899-8900: Information Technology Laboratory, NIST; October,2008

[7] K Bhaskar Prakash, Pasala Sanyasi Naidu2 "Strong and powerful cryptographic scheme for Industrial Internet of Thing using pseudo stream cipher, elliptic curve cryptography and trigonometric techniques (An efficient scheme to detect COVID-19 patient from Quarantine people) Solid State Technology Volume: 63 Issue: 6 Year: 2020

[8] Kahate Atul. Cryptography and network security. Tata McGraw- Hill; 2006.

[9] Swayam Course in Information Security(Swayam is a program initiated by Government of India and designed to achieve the three cardinal principles of Education policy viz, access, equity and quality)

[10] NIST. Secure Hash Standars. FIPS PUB; 2002. pp. 180 - 2.

[11] Preenel B. Cryptographic hash functions. Transactions on Telecommunications 1994;5:431 - 48.

[12] Rivest RL. The MD4 Message Digest Algorithm. RFC 1320; 1992.

[13] Rivest RL. The MD5 Message Digest Algorithm. RFC 1321; 1992.

[14] Rivest RL. The MD2 Message Digest Algorithm. RFC 1319; 1992.

[15] Stallings William. Cryptography and network security: Principles and practice. 3/e PH; 2003.

[16] Status Report on the First Round of the SHA-3 Cryptographic Hash Algorithm Competition; September 2009. http://csrc.nist.gov/ publications/nistir/ir7620/nistir_7620.pdf.

[17] Status Report on the Second Round of the SHA-3 Cryptographic Hash Algorithm Competition.http://csrc.nist.gov/publications/nistir/ir7764/ nistir-7764.pdf; February 2011.

[18] Third-Round Report of the SHA-3 Cryptographic Hash Algorithm Competition.http://dx.doi.org/10.6028/NIST.IR.7896; November 2012.

[19] Wang X, Yu H. How to break MD5 and Other hash functions. In: Cramer R, editor. EUROCRYPT 2005, LNCS 3494; 2005. pp. 19 - 35.

[20] Wang X, Feng XD, Lai X, Yu H. Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD. rump session, CRYPTO 04; 2004.

[21] Wang X, Yin Y, Yu H. Finding Collisions in the Full SHA-1. Lecture notes in Computer Science. CRYPTO 2005 Proceedings, Vol. 3621; 2005. pp. 17 - 36.

[22] Wang X, Feng D, Lai X, Yu H. Collisions for hash functions MD4, MD5, HAVAL-128 and RIPEMD. Cryptology ePrint archive; 2004.

Heart Diseases Prediction for Optimization based Feature Selection and Classification using Machine Learning Methods

N. Rajinikanth[1], Dr. L. Pavithra[2]
Associate Professor, CMS College of Science and Commerce
Associate Professor, Department of Computer Science, Dr. N. G. P. Arts and Science College, Coimbatore, india

***Abstract*—Globally, heart disease is considered to be the major cause of death. As per statistics, 17.9 million people are losing their lives every year worldwide. Chronic Kidney Disease (CKD) and Breast Cancer takes the next positions in the list. Disease classification is an important issue that needs more attention now. Making use of an optimized technique for such classification would be a better option. In this heart disease classification, initially, feature selection was done using Teaching learning based Optimization based (TLO) and Kernel Density. TLO is based on the process of classroom teaching, which involves too much iteration that leads to time complexity. Similarly, a certain level of misclassifications has been observed by using Kernel Density (KD). In the proposed method, K-Nearest Neighbour (KNN) is used to address the issue of NaN values and Density based Modified Teaching Learning based Optimization (DMTLO) is used for feature selection. Finally the classification process is done by considering Support Vector Machine (SVM) and Ensemble (Adaboosting method). SVM categorizes data bydissimilar class names by defining a group of support vectors that are part of the group of training inputs that plan a hyper plane in the attribute space. Ensemble method is used to solve statistical, computational and representational problems. Experimental outcomes have proved that the projected DMTLOovertakes the existing methodologies with required quantity of attributes.**

Keywords—Teaching learning based optimization; kernel density; support vector machine; k-nearest neighbour; ensemble learning

I. Introduction

Nowadays, datasets are tremendously accumulated with enormous quantity of data sources. Such high dimensional data rises the calculationrate and diminishes the results of a ML model if the dataset has inappropriate, duplicate and unwanted attributeswhich is notfavourable to the improvement of an analytical model. The issue of over fitting with vast number of features could be addressed by using Learning models. Choosing a relevant and suitable set of features could be a better way to solve this problem. Several feature selection algorithms are available in this regard. These algorithms are capable of minimizing the quantity of features in order to develop an AI model by authenticating different arrangements of features in an input dataset.

In general, wrapper based attribute selection strategies are projected to improve the competencies of classification methods. Finding a worthy arrangement of attributes is really a challenging task. Various optimization techniques are utilized for choosing proper features such as Genetic Algorithm (GA), and Particle Swarm Optimization (PSO) by numerousscientists to advance the outcomes of the classifiers.

Parham et al., (2016) [9] established an attribute choosing strategy which is a hybridization of PSO and local search strategy. Its results were evaluated with various screen and wrapper-based strategies. It has attained notable precision results.

Hafez et al. (2015) [5] proposed an attribute choosing procedure that is dependent on Chicken swarm optimization. It replicated the performance of chicken swarms and attainedgood resultsthroughtypical datasets relatedtowards GA and PSO optimization algorithms. A methodology proposed by Panda (2017) [12] relies on elephant search optimization in aalliance with deep NN for inspecting microarray data. Venkata Rao (2016) [14], Rao (2016) [21] proposed extensive presentations of TLBO in many real time problems. The strategy of TLBO is proposed to decrease load of fixing the parameter standards during attribute choosing process.

II. Related Work

Attribute selection is highly needed in various areas like categorization of emails, disease analysis, forged claims and also in the areas of credit/debit risks. In the process of developing a well-organized decision-making method, the significant step is to organize the better features which are more suitable to attain better precision results.Various scientists have made use of filter and wrapper choosing strategies Wah et al., (2018) [22] to increase the correctness of forecaststrategies. Several prevailing attribute choosing strategies have been observed to comprehend its pros and cons. Bahassine et al. (2018) [3] have projected a novel attribute choosing method for categorization of Arabic text by means of an better Chi-square technique to improve the classification outcomes. Better results have been attained by incorporating SVM classifier.

Mazini et al. (2018) [11] established a new method intended for abnormality network-based intrusion discovery model. This helps to attain a maximum detection rate with a minimum false positive rate. This model is a hybridization of both artificial bee colony and AdaBoost algorithm. The former is utilized for selecting efficient attribute whereas the latter is for classification.

Thawkar et al., (2018) [18] projected an attribute choosing method. This method was developed using Biogeography-based optimization procedureaimed atcategorization of numeral mammograms with ANN.

Wen et al. (2016) [23] developed a novel unsupervised attribute choosing techniquethat is related on L2;1-norm regularization on behalf ofidentifying certain human movements. The above said procedure achieves both attribute mining and selection instantaneously which produces ideal attributes.

Xu et al. (2017) [24] projected an innovative discriminative L2 regularization-based sparse demonstration. This procedure is exclusively for classifying input images and accomplishednotableprecisionthroughvarious inputs.

Absolute dimensionality reducing method is proposed by Lai et al. (2017) [7] that can be termed asRobust Discriminant Regression (RDR) by means of L2,1-norm as the elementarystandard in the evaluation function for attribute extraction. RDR doesn't get proper predictions for attribute selection and that is considered to be its main disadvantage.

Mafarja et al. (2017) [8] utilized the Dual Dragonfly Procedure. This is in the direction ofpicking a subdivision of attributes taken fromUCI repository and attainedimprovedout comesequated with GA and PSO algorithms.

Sayed et al. (2017) [15] recommended a fresh meta-heuristic techniquewhich is similar to crow search procedure for picking proper attributes and appealedhealthieroutcomes throughstandard datasets.

Sayed et al. (2018) [16] established a hybridized technique which is a combination of swarm algorithm for attribute selection and with chaos theory. This addresses the issues of confined optima and little convergence problems.

Agrawal et al., (2015) [1] projected a novel attribute selection strategy that is dependent on Artificial Bee Colony and K-NN algorithms. This is used for categorizing theCT images of cervical cancer.

Marie-Sainte et al., (2018) [10] recommended an innovative attribute choosing method for categorizing Arabic text with the help of firefly algorithm. This obviously improves classification performance. The researchershave made trials on OSAC dataset and accomplished 0.994 accuracy rate.

Shahbeig et al. (2016) [17] designated a subcategory of interrelatedDNAscollected from the input of breast cancer microarraythrough the support of transformed fuzzy adaptive PSO incorporated with TLBO procedure and confirmed the correctnessby SVM classifier.

Tuo et al. (2017) [19] established an original hybrid HSTLBO technique that stabilizes the convergence difficulty of distinct TLBO and Harmony Search procedures.

III. Existing System

Feature selection can be done in dualways; Teaching Learning based Optimization (TLO) and Kernel Density (KD)

A. Feature Selection using Teaching Learning based Optimization (TLO)

TLO is familiar technique towards choosing the ideal sub division of features. This has binary segments. First segment covers an optimization Technique, which can be utilized to choose ideal set of attributes. Various classification models are covered in the latter phase. These segments arerecurrenttill anendingcondition has seen. Stopping criteria can be taken as astaticamount of iterations. Improved precision with various classification models cannot be adopted in Teaching Learning based Optimization (TLO) and also this TLO cannot be hybridized with any other feature selection strategies.

B. Feature Selection using Kernel Density (KD)

Kernel Density (KD) isa non-parametric and it doesn't make any conventions with respect to data distribution. It always chooses attributes that capture the performance of usual data by separating the outliers. A forward search strategy is used for estimating standards. This is highly capable of discovering outliers when compared to other familiar strategies. Incorporating other search techniques would be a more challenging factor in terms of attribute selection since it exploits the parallelism. Also, no proper studies have been done so far to ensure the value of the features.

IV. Proposed System

In the proposed system pre-processing to remove the Nan is done using KNN method, feature selection using Density based Modified Teaching Learning based Optimization (DMTLO) and Kernel Density (KD) based method. Classification is done using Classification using SVM and Ensemble (Adaboosting method).

Fig. 1 represents the architecture of the proposed system.

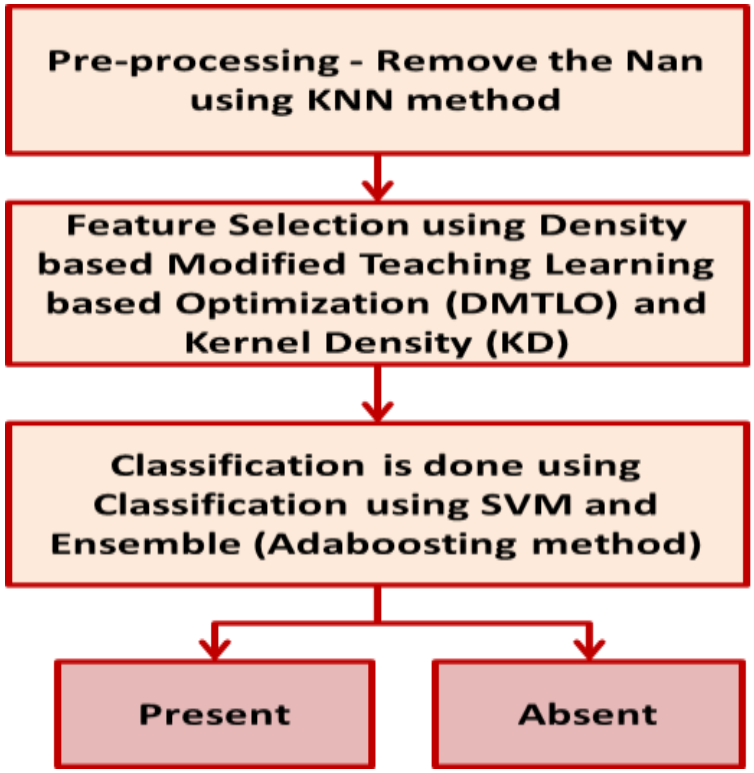


Fig. 1. Architecture of the Proposed System

A. *Pre-Processing to Remove the Nan using KNN Method*

In familiar data mining tasks like, classification and regression Altman (1992) [2], K-Nearest Neighbour (K-NN) is considered to be a constraintfree approach. It is a method of instance-based learning and it is likewise termed as lazy learning. Local approximations are done on the functions and the calculations are suspended until classification. It is considered to be the basic way of all AI techniques.

Its outcomes determine the classification or regression. The characteristicscomprise ease to take outcomes, calculation time and analyticalcompetence.If K-NN is utilized for classification, the results give the class membership.

Objects are categorizedby means ofconsidering the vote attained from neighbours. All those objects are allocated to a class which is more obvious in KNN.In the phase of regression, the outcome provides the stuff of object which is the average of the values of KNNs.

B. *Feature Selection using Density based Modified Teaching Learning based Optimization (DMTLO)*

Density based Modified Teaching Learning based Optimization (DMTLO) is adopted inorder to streamline the conventional TLBO in the calculation of evaluation function. The size of input and design variables is considered to be theinput parameters to discover the biasedgroup of attributes.

DMTLO starts by fixing the populacescope, t i.e., the quantity of learners (ps= 1, 2, 3 ….. n) and the design variable, s i.e., the quantity of subjects (su=1, 2, 3…..m) which aretrained.

The representation of objective function is given below.

$$\text{Minimum } f(y) = \sum nr=1 \, [y2\, r - 10 \cos(2\pi yr) + 10] \quad (1)$$

1) Teacher phase: The best learner would be chosen in this phase. Teacher tries to take an attempt in order to enrich the understanding of rest of the learners by maximizing their average mean. Throughout this phase, final iteration can be represented as.

y^{th}Iteration for (y=1, 2, 3….m)

Subject x(x=1, 2, 3…..n)

Mean value for individual subject is considered and it could be demonstrated as ms(x,y)

In this phase, variancesare taken to modernize the standards in the resolutionpool by totalling the value of differences to the present solution and the algorithm continues to the learner phase.

Chebyshev distance metric is taken to modernize the values in output space. Differences are denoted as Ds, Dchebyshev distance as Dc.

$$Ds = v\,(O\_new,s - TFOs) \quad (2)$$

$$Dc\,(yi, yj) = \max(|yi - yj|) \quad (3)$$

$$X'new = f(y) + Dc(yi - yj) \quad (4)$$

2) Learner phase: By making interaction with the peers, the understandability of individual learners can be improved.

For y=1:tr

Chooseadditional learner arbitrarily Xx, such that y≠x

If f(Xx) < f(Xx)

X"new,y=X'new,y + ry(Xy– Xx)

Else

X"new,y=X'new,y+ ry(Xx – Xy)

End If

End For

Admit‘X"new’, when a function value is superiortoits earlier value. The attributes that shows enhanced outcomes based on the latest evaluation function through the every cycle is accumulated in attribute subset. This algorithm finishes when each and every attributes are taken for evaluation.

C. Classification using SVM and Ensemble (Adaboostingmethod)

Classification is done using Classification using SVM and Ensemble (Adaboosting method).

1) Support Vector Machine (SVM): One of the newestproceduresaimed at pattern classification is SVM. It isextensively used in various fields. It is a supervised learning techniqueconnected with learning procedures to examine data and to distinguish patterns. Fixing up the kernel factor for SVM in training phase will definitely influence the correctness of classification results. SVMs were initially recommended by Vapnik (1995) [20]. It is widely used in various applications like image recognition Pontil&Verri (1998) [13], bioinformatics Yu et al. (2003) [25] and text classification Joachims (1998) [6].

Class labels are used to classify the input data. This is possible via defining a group of support vectors which areconsidered to be a part of training inputs.

Along with linear classification, SVMs are well relevant for random classification with the help of data, indirectlyplotting their inputs on high-dimensional attribute spaces.

2) Ensemble classification: Ensemble learning helps in enlightening the outcomes of Machine Learning (ML) by linking several models. This strategy produces a notable outcome in contrast to a solitary model. A group of classifiers acquire and then cast their vote. The extrapolativecorrectness is upgraded but it is challenging to comprehend them Dietterich (2002) [4]. It is beneficial in solving statistical, computational and representational problems. It is not essential to find more precise models, but build models with errors. Ensemble models built to perform classification can misclassify initially.

There are different methods of building ensembles.

- Maximum Vote
- Bagging and Random Forest (RF)
- Chance Injection
- Feature choice Ensembles
- Error Correcting Output Coding (ECOC)

The algorithm is shown below.

Step 1: Form the test set ‘T’ using ‘n’ documents in ‘X’

Step 2: Form the training set ‘TR’ using the residual documents in ‘X’

Step 3: for every classifier in ‘C’.

Make use of classified documents to train the classifier in ‘T’.

Utilize the trained classifier to group the documents in ‘S’.

Store the resultant labels in the particular class.

Step 4: for every ‘x’ in the range 1 to s

for every ‘y’ in the range 1 to s

for every ‘z’ in the range 1 to k

for every ‘n’ in the range z+1 to k

if (class[z,x] == class[n,y])

if(M[x,y]==0)

M[x,y]=1;

else

M[x, y] =M[x, y]*2;

Step 5: ‘m’ is served into the k-means procedure to form document groups.

Step 6: Apply SVM-linear algorithm on ‘T’ for document categorization.

Step 7: Select the classes conforming to the clusters by finding the class attained in the preceding step.

D. Datasets

The datasets aretaken from UCI machine learning repository.

Nearly 76 features are present in the heart disease dataset, but most of the researchers have made use of 14 in the list. The objective of this dataset is to conclude whether a patient is having a heart disease or not. It is numericalvalue that ranges from 0 to 4. Investigations with the Cleveland database have focused on simply attempting to differentiate existence (values 1, 2, 3, 4) from non-existence (value 0) of heart disease.

In the heart diseases dataset there are 14 attributes 304 Instances, whereas in Chronic Kidney Disease dataset there are 25 attributes 400 Instances and Breast cancer dataset includes 32 attributes 569 Instances. Each has an attribute that is a class like present and not present.

Chronic Kidney Disease dataset includes blood tests and various other measures collected from the patients either with the presence or absence of CKD. The details are collected from nearly 400 patients who were in observation for over period of 60 days. Out of 400 patients, 250 were diagnosed with Chronic

Kidney Disease and 150 were without Chronic Kidney Disease. This variation is represented as "Class" in the dataset. Few important attributes of this dataset are age, Hyper tension, Diabetic, Blood Glucose Random, Blood Urea, Haemoglobin etc.

Wisconsin Diagnostic Breast Cancer (WDBC) is one of the standard datasets considered for Breast cancer diagnosis. It has nearly 699 instances, in which 458 are benign and 241 are malignant with 11 attributes that includes a class attribute.

V. Results and Discussion

The following figures (Fig. 2-7) show the performance of the benchmarked and the proposed schemes. Table I shows the quantity of Features taken using TLO, KD and DMTLO. Table II illustrates the attributes selected using TLO, KD and DMTLO.

Fig. 2 shows the Accuracy, Precision, Recall, F-measure for the Heart Disease Dataset. It is seen that the proposed DMTLO_Adaboosting offers 6%, 4%, 4%, 2%, and 2% better Accuracy in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboostingand DMTLO_SVM respectively. Similarly it offers 5%, 3%, 3%, 1%, and 1% better Precision in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively. The Recall of DMTLO_Adaboosting is 4%, 4%, 1%, 2% and 2% improved when compared to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM respectively. Similarly, the F-Measure of DMTLO_Adaboosting is 4%, 3%, 3%, 2% and 1% improved when compared to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively.

Fig. 3 shows the Time Period and Error rate for the Heart Disease Dataset. DMTLO_Adaboosting offers 80.95%, 61.90%, 61.90%, 33.33% and 23.81% better Time period in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM respectively. Similarly it involves 2.76, 2.46, 2.23, 1.85 and 1.38 times lesser error rate in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively.

Fig. 4 shows the Accuracy, Precision, Recall, F-measure for the Chronic Kidney Disease Dataset. It is seen that the proposed DMTLO_Adaboosting offers 6%, 3%, 7%, 4%, and 2% better Accuracy in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM respectively. Similarly it offers 6%, 3%, 7%, 5%, and 2% better Precision in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively. The Recall of DMTLO_Adaboosting is 5%, 3%, 7%, 4% and 2% improved when compared to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM respectively. Similarly, the F-Measure of DMTLO_Adaboosting is 5%, 3%, 6%, 3% and 1% improved when compared to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively.

Fig. 5 shows the Time Period and Error rate for the Chronic Kidney Disease Dataset. DMTLO_Adaboosting offers 65.21%, 39.13%, 86.95%, 60.86% and 26.08% better Time period in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively. Similarly it involves 2, 1.58,2, 1.75 and 1.33 times lesser error rate in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively.

Fig. 6 shows the Accuracy, Precision, Recall, F-measure for the Breast Cancer Dataset. It is seen that the proposed DMTLO_Adaboosting offers 5%, 1%, 5%, 3%, and 2% better Accuracy in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively. Similarly it offers 5%, 2%, 4%, 3%, and 1% better Precision in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively. The Recall of DMTLO_Adaboosting is 5%, 2%, 5%, 4% and 1% improved when compared to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively. Similarly, the F-Measure of DMTLO_Adaboosting is 3%, 1%, 6%, 1% and 1% improved when compared to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively.

TABLE I. Number of Features Selected using TLO, KD and DMTLO

Dataset	No. of Attribute Selection	Feature Selection using TLO	Feature selection using KD	Feature Selection using DMTLO
Heart disease	14	10	8	12
CKD	25	18	16	19
Breast Cancer	32	20	23	26

TABLE II. Attributes Selected using TLO, KD and DMTLO

Dataset	Selected Attributes of TLO	Selected Attributes of KD	Selected Attributes of DMTLO
Heart diseases	1,2,3,4,5,6,7,10,12,13	4,5,6,12,13,10,7,3,	12,10,11,8,13,2,4,7,6,5,9
CKD	2,3,4,5,6,10,17,18,19,14,15,11,13,12,9,8 ,16,20	3,4,5,10,11,12,15,16,19,18,11,8,9,2,14,6	2,3,10,4,5,17,18,19,14,15,6,7,11,12,13,8,9,21,23
Breast Cancer	12,13,11,27,28,8,7,29,6,18,17,16,19,10, 15,14,22,21,26	11,12,13,14,17,27,28,29,30,15,16,17,23,22,2 1,18,19,2,3,4,5,24,8,	12,13,11,27,28,29,26,8,7,25,9,5,18,30,17,16,19, 10,15,2,14,1,22,21,6,4

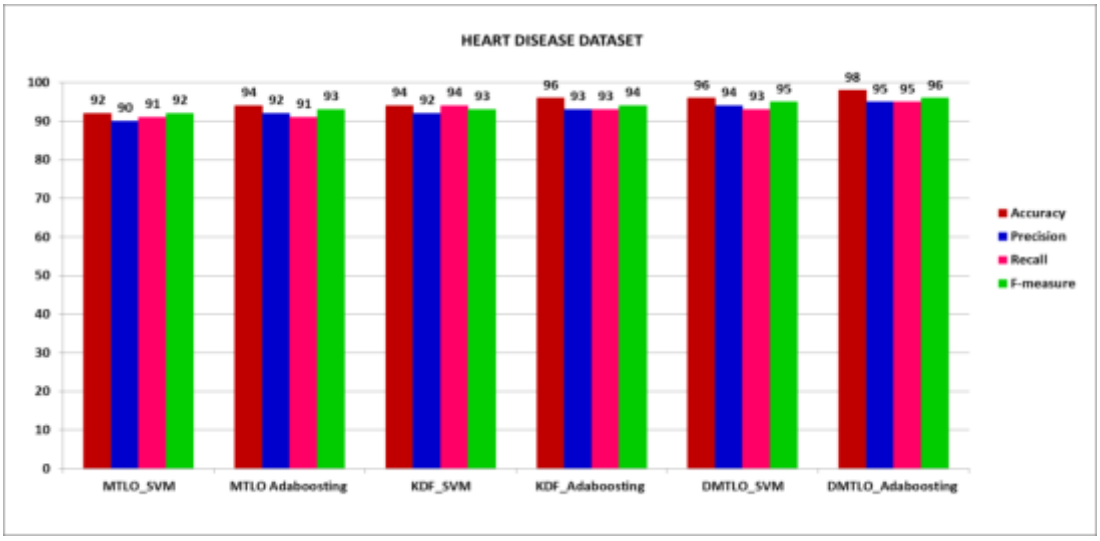


Fig. 2. Accuracy, Precision, Recall, F-Measure for the Heart Disease Dataset.

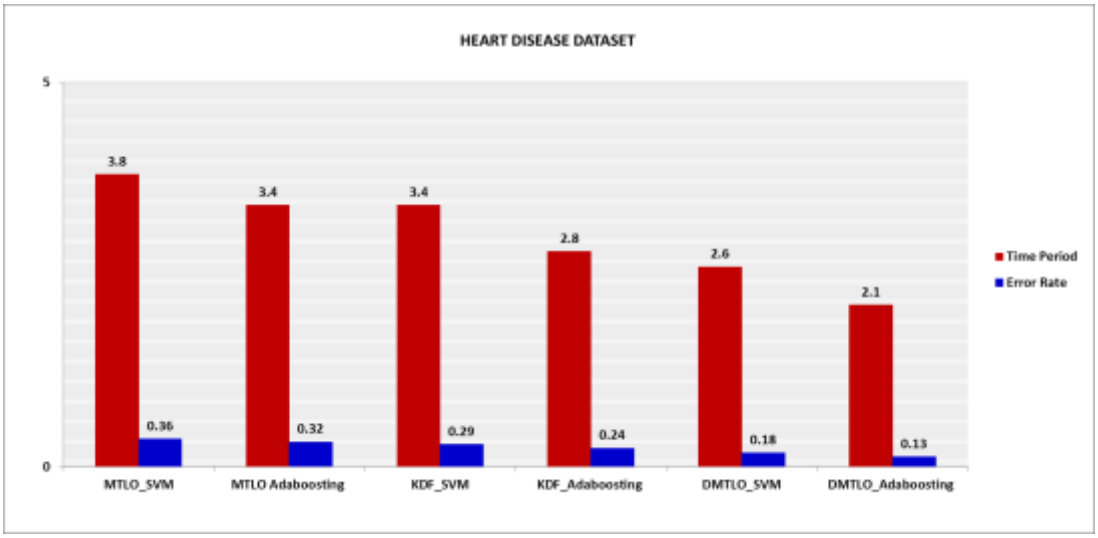


Fig. 3. Time Period and Error Rate for the Heart Disease Dataset.

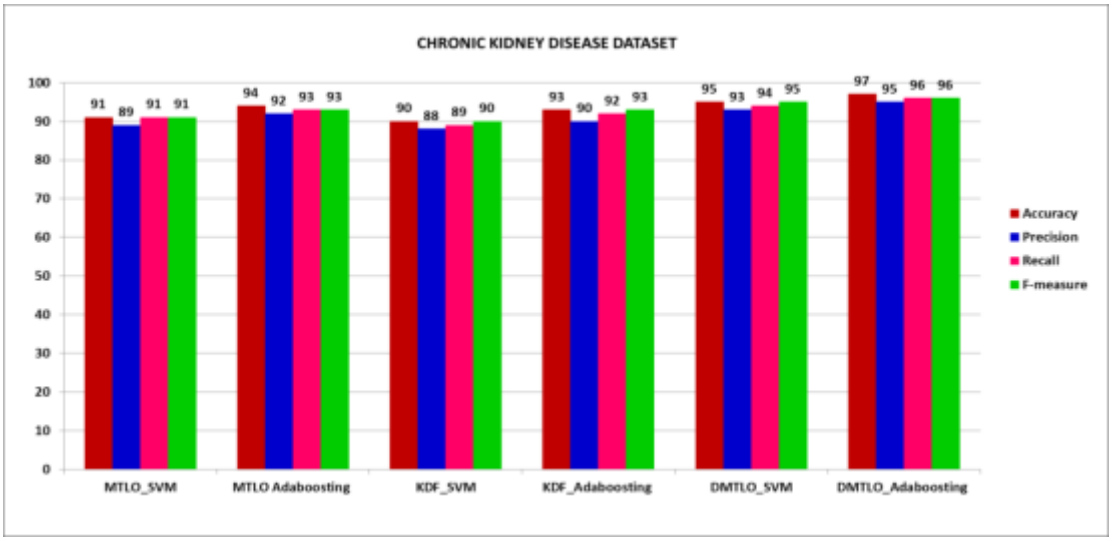


Fig. 4. Accuracy, Precision, Recall, F-Measure for the Chronic Kidney Disease Dataset.

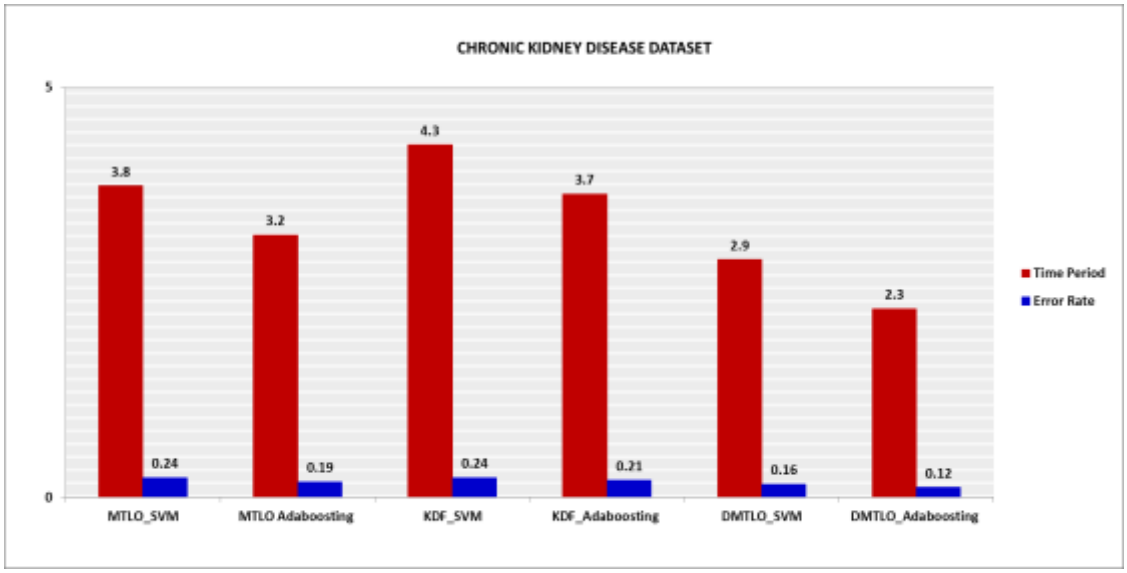


Fig. 5. Time Period and Error Rate for the Chronic Kidney Disease Dataset.

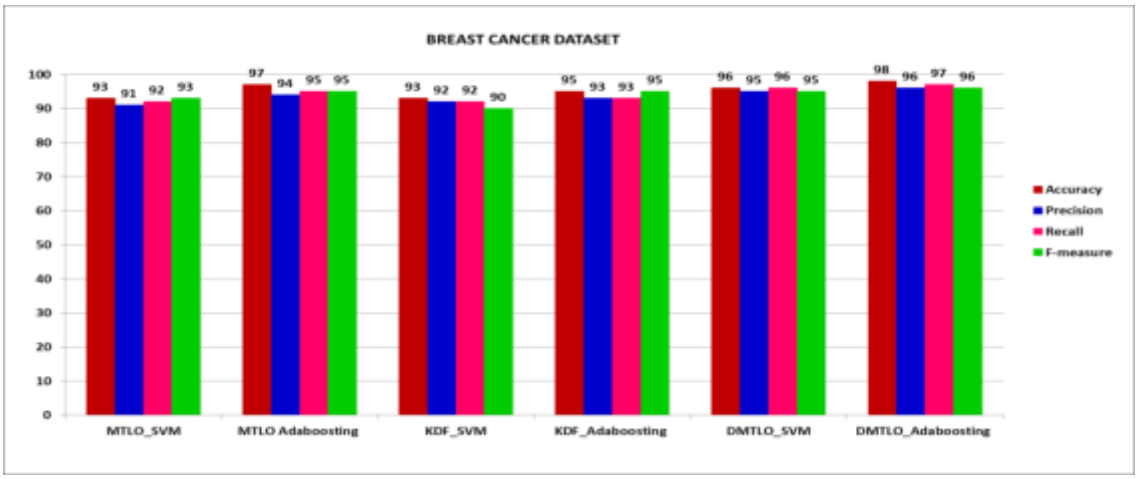


Fig. 6. Accuracy, Precision, Recall, F-Measure for the Breast Cancer Dataset.

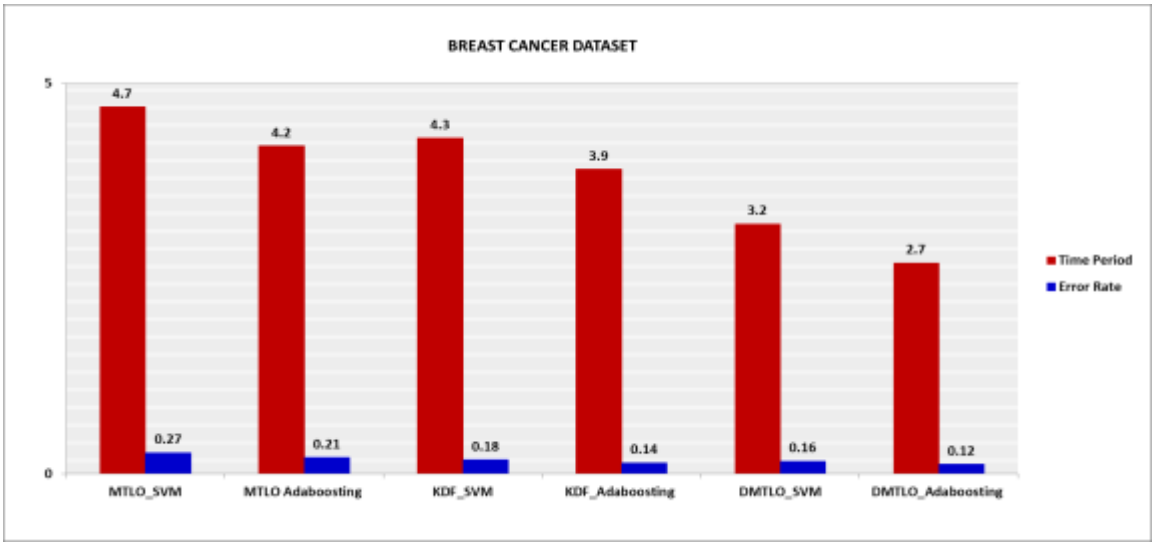


Fig. 7. Time Period and Error Rate for the Breast Cancer Dataset.

Fig. 7 shows the Time Period and Error rate for the Breast Cancer Dataset. DMTLO_Adaboosting offers 74.07%, 55.55%, 59.25%, 44.44% and 18.51% better Time period in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively. Similarly it involves 2.25, 1.75, 1.5, 1.66 and 1.33 times lesser error rate in contrast to MTLO_SVM, MTLO Adaboosting, KDF_SVM, KDF_Adaboosting and DMTLO_SVM, respectively.

VI. Conclusion

In this paper, the outcomes of the proposed system are evaluated for 3 various datasets like Heat disease, chronic kidney disease and Breast cancer. The experimental results are compared with existing Teaching Learning optimization and Kernel Density. The results are analysed in terms of Accuracy, Precision, Recall, F-measure, Time Period and Error Rate. Based on this, it is noticeable that the proposed DMLTO overtakes the existing methodologies.

References

[1] Agrawal, Vartika, Chandra, Satish, 2015. ‘‘Feature Selection using Artificial Bee Colony Algorithm for Medical Image Classification”. In: International Conference on Contemporary Computing (IC3). IEEE, pp. 171–176.

[2] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.

[3] Bahassine, Said, Madani, Abdellah, Al-Sarem, Mohammed, Kissi, Mohamed, 2018. Feature selection using an improved Chi-square for Arabic text classification. J.King Saud Univ. – Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2018.05.010.

[4] Dietterich, T. G., 2002, “Ensemble learning, The handbook of brain theory and neural networks”, MA Arbib, vol. 2, pp. 110-125.

[5] Hafez, A.I., Zawbaa, H.M., Emery, E., Mahmoud, H.A., Hassanien, A.E., 2015. ‘‘An innovative approach for feature selection based on chicken swarm optimization”. 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), pp. 19–24.

[6] Joachims, T 1998, ‘Text categorization with support vector machines: Learning with many relevant features’, Springer Berlin Heidelberg, pp. 137-142.

[7] Lai, Zhihui, Mo, Dongmei, Wong, Wai Keung, Yong, Xu., Miao, Duoqian, Zhang, David, 2017. ‘‘Robust Discriminant Regression for Feature Extraction”. IEEE Trans. Cybernetics. 10.1109/TCYB. 2017.2740949.

[8] Mafarja, M.M., Eleyan, D., Jaber, I., Hammouri, A., Mirjalili, S., 2017. ‘‘Binary Dragonfly Algorithm for Feature Selection,”. International Conference on New Trends in Computing Sciences (ICTCS), pp. 12–17.

[9] Moradi, Parham, Gholampour, Mozhgan, 2016. ‘‘A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy”. Appl. Soft Comput. 43, 117–130.

[10] Marie-Sainte, Larabi, Alalyani, S., 2018. (in press), N. Firefly Algorithm based Feature Selection for Arabic Text Classification. J. King Saud Univ. – Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2018.06.004.

[11] Mazini, Mehrnaz, Shirazi, BabakIrajMahdavi, 2018. (in press), ‘‘Anomaly network based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms”. J. King Saud University – Comput. Inf. Sci. https://doi. org/10.1016/j.jksuci. 2018.03.011.

[12] Panda, M., 2017. Elephant search optimization combined with the deep neural network for microarray data analysis. J. King Saud Univ. – Computer Inf. Sci. https://doi.org/10.1016/j.jksuci.2017.12.002.

[13] Pontil, M &Verri, A 1998, ‘Support vector machines for 3D object recognition’, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 6, pp. 637-646.

[14] Rao, R.V., 2016. Teaching Learning Based Optimization Algorithm And Its Engineering Applications. Springer International Publishing.

[15] Sayed, G.I., Hassanien, A.E., Azar, A.T., 2017. Feature selection via a novel chaotic crow search algorithm. Neural Comput. Appl.

[16] Sayed, G.I., Khoriba, G., Haggag, M.H., 2018. ‘‘A novel chaotic salp swarm algorithm for global optimization and feature selection”. Appl. Intell., 1–20.

[17] Shahbeig, Saleh, SadeghHelfroush, Mohammad, Rahideh, Akbar, 2016. ‘‘A Fuzzy Multi-Objective Hybrid TLBO-PSO Approach to Select the Associated Genes with Breast Cancer”. Signal Process. 131, 58–65.

[18] Thawkar, Shankar, Ingolikar, Ranjana, 2018. (in press), ‘‘Classification of masses in digital mammograms using Biogeography-based optimization technique”. J.King Saud Univ. – Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2018.01.004.

[19] Tuo, Shouheng, Yong, Longquan, Deng, Fang’an, Li, Yanhai, Lin, Yong, Qiuju, Lu,2017. ‘‘HSTLBO: a hybrid algorithm based on Harmony Search and Teaching- LearningBased Optimization for complex highdimensional optimization problems”. PLoS One 12.

[20] Vapnik V 1995, ‘Support-vector networks’, Machine Learning, vol.20, pp. 273–297.

[21] Venkata Rao, R., 2016. Review of applications of TLBO algorithm and a tutorial for beginners to solve the unconstrained and constrained optimization problems. Decision Sci. Lett. 5, 1–30.

[22] Wah, Y.B., Ibrahim, N., Hamid, H.A., Abdul-Rahman, S., Fong, S., 2018. Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. Pertanika J. Sci. Technol. 26 (1), 329–340.

[23] Wen, Jiajun, Lai, Zhihui, Zhan, Yinwei, Cui, Jinrong, 2016. The L2,1-norm-based unsupervised optimal feature selection with applications to action recognition.PatternRecogn. https://doi.org/10.1016/j.patcog.2016. 06.006.

[24] Xu, Y., Zhong, Z., Yang, J., You, J., Zhang, D., 2017. A new discriminative sparse representation method for robust face recognition via L2 regularization. IEEE Trans. Neural Networks Learn. Syst. https://doi.org/10.1109/TNNLS.2016.2580572.

[25] Yu, GX, Ostrouchov, G, Geist, A &Samatova, NF 2003, ‘An SVM-based algorithm for identification of photosynthesis-specific genome features’, In Bioinformatics Conference Proceedings of the IEEE, pp. 235-243.

Face Recognition based on Convolution Neural Network and Scale Invariant Feature Transform

Jamilah ALAMRI[1], Rafika HARRABI[2], Slim BEN CHAABANE[3]
Faculty of Information Technology, Department of Information Technology
Industrial Innovation and Robotics Center
University of Tabuk, Tabuk
Kingdom Saudi Arabia

***Abstract*—Recently, Face Recognition (FR) has been received wide attention from both the research community and the cyber security industrial companies. Low accuracy of recognition is considered a main challenge when it comes to talking about employing the Artificial Intelligence (AI) for FR. In this work, the Scale Invariant Feature Transform (SIFT) and the Convolutional Neural Networks (CNN) feature extraction methods are utilized to build an AI based classifier. The CNN extracts features through both the convolutional and polling layers, while the SIFT extracts features depending on the scale space, directions, and histograms of points of interest. The features that are extracted by the CNN and the SIFT methods are used as an inputs for the KNN classifier. The experimental results with 400 test images of 40 persons, with 240 images are randomly chosen as training sets and 160 images from test sets, demonstrate in terms of accuracy, sensitivity, and error rate, that the CNN-based KNN classifier achieved better results when compared to the SIFT-based KNN classifier (accuracy = 97%, sensitivity = 93%, error rate = 3%).**

Keywords*—Face recognition; training; testing; CNN; SIFT; accuracy; classifier

I. Introduction

Background The Face Recognition (FR) research field can be seen as an intersection of three main domains, which are Artificial Intelligence (AI), Image Processing (IP), and Cybersecurity (Cs). Fig. 1 illustrates the face recognition research field in terms of domains' intersection.

For the Artificial Intelligence (AI), it is defined as the science that addresses the mechanisms of learning machines to be able to make decisions as the human's brain [1]. The Image Processing (IP) research field is defined as method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it [2]. Cybersecurity is defined as the mechanisms that are employed to protect digital data against unauthorized network users or malicious alternations [3].

Motivation (importance of domain). In the context of Smart Cities (SCs), FR-based systems play a significant role to perform tasks easily and quickly for the users. FR-based systems can save the user's time. For example, instead of opining door using keys, the FR-based system can do this mission directly once the user stands in front of the door. This saves the time of the user when he or she forgets the keys of the door [4]. Moreover, FR-based systems ensure performing the tasks at a high level of security. That is because nobody can login to sensitive locations (or data) if the system denies the matching process [5]. Furthermore, from medical point of view, FR-based systems contribute to limit the spread of Covid-19. That is because fingerprinting-based systems can be replaced by FR-based systems [6]. Actually, it is recommended not to use fingerprinting systems in both governmental and private institutions (PMC, 2020).

Statement of problem. In terms of cybersecurity, authentication security requirement means the process of identifying the identity of a user with guaranteeing that no impersonation [7]. FR contributes to provide authentication for users by processing the image of the user's face and then matching it with what was stored in a database. However, employing the Artificial Intelligence (AI) to build FR-based system is critical especially when it comes to talking about logging in to a top-secret data centers, such as the servers' room in an interior ministry, or any kind of digital information [8]. That is because any error in the FR-based system leads to open the door for attackers (unauthenticated users) [9]. This in turn means a very critical security gap in the system.

This reflects the importance of providing FR-based systems with a high accuracy. Otherwise, a big security problem will occur. Fig. 2 illustrates the problem of low accuracy of FR-based systems from cybersecurity perspective.

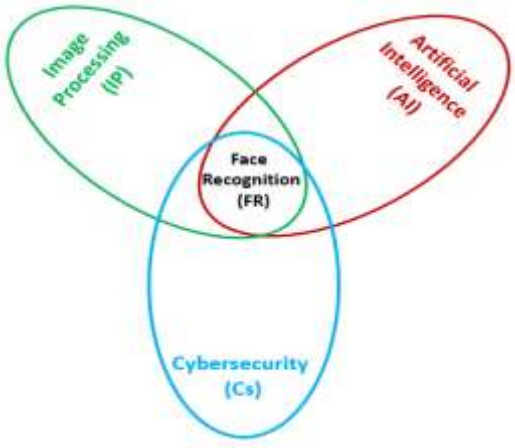


Fig. 1. Face Recognition Research Field in Terms of Domains' Intersection.

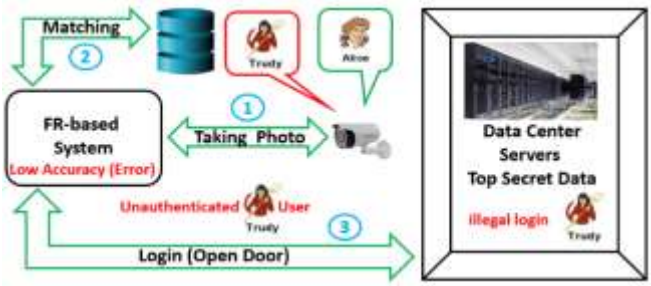


Fig. 2. The Problem of Low Accuracy of FR-based Systems.

In Fig. 2, there are two users (Alice and Trudy). Alice is authenticated user, while the Trudy is un-authenticated one. The FR-based system is linked to a data center that includes servers where a top-secret data is stored in it. Three main steps are required to legal login to the data center. First, the camera takes a photo of the face of the user. Second, the FR-based system processes the image and performs a matching process. If the information extracted from the processed image (the face of the user is recognized) matches with what is stored in the data base, the login is performed (physically, the door is opened). In the case of low accuracy, the FR-based system will have a security gap. This security gap can be exploited by the Trudy to gain an illegal login.

Research questions. There are some critical reasons of the low accuracy in FR-based systems. All of the reasons have a one root related to blurring. From the term of blurring, a main research question can be derived, which is how to ensure robustness against blurring images of the faces and high accuracy of face recognition at the same time? In details, we have the following tow research questions:

1) How to ensure high face recognition rate under the impact of noisy images of faces (such as wet faces or sweaty face).

2) How to guarantee high accuracy when dealing with different cases where the directions of the faces cause some distortion of the face.

By employing Convolution Neural Networks (CNN), we can response to the research questions. We can exploit the structure of the CNN that contains constructing the convolution and pooling layers to enhance the processing of the input images. In addition, we can support the CNN by a strong pre-processing step to ensure high resistance against noisy images.

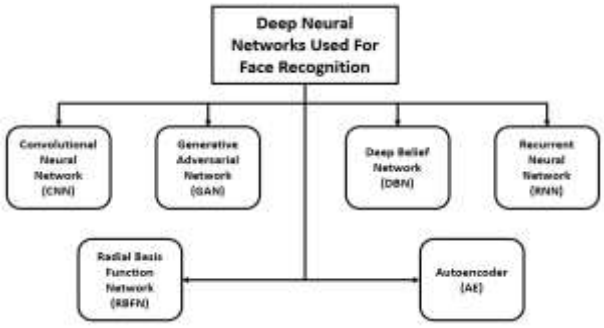


Fig. 3. Groups of Deep Learning Networks used for FR-based Systems.

Contribution. In general, the contribution of this work is as follows:

- In responding to the first research question, an efficient pre-processing step is conducted before starting the process of training the classifier. One of the aims is removing noise.
- To deal with the distortion caused by looking at different directions, this work presents the Convolutional Neural Network (CNN) and the Scale Invariant Feature Transform (SIFT) methods for feature extraction. Integration of both the CNN and the SIFT with the K-Nearest Neighbor (KNN) classifier ensure high level of accuracy under various directions of the face being classified.
- Extensive experiments are conducted to proof the effectiveness of the proposed classifiers.

Structure of paper. The rest of the work is organized so that in Section II we present the related work. Section III provides the proposed system. In Sections IV and V, the used metrics and the experiments and evaluations are conducted, respectively. Finally, the work is concluded in Section VI.

II. Related Work

In general, the Artificial Intelligence (AI) provides significant contribution to enhance the FR-based systems. Under the umbrella of the Artificial Intelligence (AI), the deep learning networks that are used for building FR-based systems can be classified into six groups, as shown in Fig. 3.

In [10] a suggested algorithm was proposed to increase the efficiency of the Elman neural algorithm in face recognition. The proposed algorithm was studied on the images of 20 students from the Department of Computer Science, Tikrit University. First step creates dataset of faces, second step convert color space to HSI and using saturation layer, image decomposition using curve let transform, feature extraction using Principle component analysis, and final step face recognition using Elman neural network. After applying proposed algorithm, the rate of face recognition 94%."

In their work [11], the authors proposes an algorithm for face detection and recognition based on convolution neural networks (CNN), which outperform the traditional techniques. In order to validate the efficiency of the proposed algorithm, a smart classroom for the student's attendance using face recognition has been proposed. The face recognition system is trained on publically available labeled faces in the wild (LFW) dataset. The system can detect approximately 35 faces and recognizes 30 out of them from the single image of 40 students. The proposed system achieved 97.9% accuracy on the testing data. Moreover, generated data by smart classrooms is computed and transmitted through an IoT-based architecture using edge computing. A comparative performance study shows that our architecture outperforms in terms of data latency and real-time response.

In [12], an efficient face recognition method using AGA and ANFIS-ABC has been proposed. At first stage, the face images gathered from the database are preprocessed. At

Second stage, an interest point which is used to improve the detection rate consequently. The parameters used in the interest point determination are optimized using the Adaptive Genetic Algorithm. Finally using ANFIS, face images are classified by using extracted features. During the training process, the parameters of ANFIS are optimized using Artificial Bee Colony Algorithm (ABC) in order to improve the accuracy. The performance of the proposed ANFIS-ABC technique is evaluated using an ORL database with 400 images of 40 individuals, YALE-B database with 165 images of 15 individuals and finally with real time video the detection rate and false alarm rate is compared with proposed and existing methods to prove the system efficiency.

In [13] the authors have presented the feature-based method for 2D face images. Speeded up robust features (SURF) and scale-invariant feature transform (SIFT) are used for feature extraction. Five public datasets, namely Yale2B, Face 94, M2VTS, ORL, and FERET, are used for experimental work. Various combinations of SIFT and SURF features with two classification techniques, namely decision tree and random forest, have experimented in this work. A maximum recognition accuracy of 99.7% has been reported by the authors with a combination of SIFT (64-components) and SURF (32-components).

In [14], introduced a method to gain the invariant illumination signs of face images based on logarithmic fractal dimension with respect to complete 8-local dimensional patterns. This method depended on performing three tasks identified by using adaptive holomorphic filter to shrink the illumination partly. Second, implement the abstracted LFD method to improve facial aspects. Third, employ the full ELDP (CELDP) that utilizes the directions and the magnitude of the edge to generate the term of illumination invariant representation. The realized results based Yale B, extended Yale B and AR achieved the database results depending on their applications. The proposed method demonstrated colossal recognition excellence by reaching the entire face recognition accuracy by 99.47% for Yale B, 99.53% for CMU-PIE, 94.55% for extended Yale B, and 86.63% for AR face databases.

The author in [15] furthermore, [14] presented krawtchouk polynomial moments technique based methodology for local descriptor. Based on edge indicator, canny edge was employed to discover the focused points to specify the zone near its scale and normalize the relation. The krawtchouk polynomial will be applied on the realized region in order to construct the descriptor. The output of the ORL, FERET based method emphasized that the results were perfect confirmed by accuracy rate of (97.86) percent.

Another technique presented in [16] named Coupled Marginal Discriminant Mappings (CMDM), which matches the images of the face with different clarity levels regardless the conditions of global data distribution and local data structure based learning map. The accuracy results obtained based on AR and FERET were realized by (94.56) and (88.5) percent respectively. Additionally, dimensionality reduced local directional pattern (DRLDP) approach proposed by [17] which showed eight-bit code assigned to (3 × 3) of every sub zone. The code describes the textural pattern of the whole block and then obtains a sole eight-bit code for each block. Experiments were performed utilizing the FERET, Expanded YALE B and ORL repositories. DR-LDP beat the other local descriptor form with a higher identification score of 97.62 percent.

In addition, [18] suggested a facial recognition method focused on PCA, which was introduced using the principle of neural networks. The system's operating theory begins as follows: build a database of recognized individuals with facial images. Then agree on a training range of M number of images corresponding to the variation in facial expressions and lighting conditions of each person. Next, calculate (M × M) matrix (L) and corresponding eigenvectors and its eigenvalues. Then, fuses uniform image training set that generates M Eigen-faces and saves the corresponding values after fusing the image training set together. In addition, the program measures and stores a function vector for anyone in the database to create a different neural network designed for the face of each person found in the facial database. When Eigen-faces are collected, the corresponding computation is performed to acquire feature vectors for the facial images in the database and is given as feedback for increasing neural network training. The training method uses the facial features the same specific person's vectors that are used to train the neural network of an entity and even other neural networks. Once an input image for the identification system is provided then the resulting attribute vectors are determined using already specified Eigen-faces and the new input image representation is retrieved. The ORL face image repository has been used to test the device to demonstrate fair identification rates of 93 percent.

The author in [19] builds an automated method for identifying neutral faces in identification images utilizing deep learning algorithms, and torching the hardware necessary to efficiently execute the established environment for learning consists of 64 GB of RAM and strip-based storage unit. Free CV (open source computer vision), python and Ubuntu (Linux operating system) version 17.10 and Nvidia CUDA 8.0 are the necessary applications. This method was developed using a dataset containing approximately 94 images, the dataset was generated utilizing 128-d embedding for each face in the dataset, the embedding was used to identify the facial pictures characters. After the dataset and folder structure was set, the faces in our training set were quantified using 128 embedding. During classification, the k-NN model was utilized for the final face classification. The system achieved an accuracy of about 95%. It was able to recognize and display the names and face of people in an image. The system can recognize a face image included in a dataset that has been trained.

The study in [20] suggested a Retinex Adaptive Attenuation Quantification (AAQR) approach to improve the overnight image information. This approach contains 3 stages: the constraint of attenuation, the estimation of attenuation and the quantification of adaptations. The efficiency of the proposed model was assessed using a reliable face recognition system via sparse depiction. At night the captured driver's face images were grouped into three categories (UP-Down, Left-Right and Mixed) according to the arrangement of

illumination for each image. The findings revealed that the image recognition levels improved by the suggested AAQR system were 82%, 84%, and 91% respectively for the Up-Down, Left-Right, and Mixed Illumination classes. The detection range of the AAQR system was (2 – 36) percent higher relative to other form of picture improvement. The developed system of successful face recognition focused on the concept element interpretation, genetic algorithm and vector supporting system, in which the key aspect analysis is used to minimize the attribute aspect, the genetic algorithm is utilized to refine the searching technique, and the assistance of the vector system is employed to recognize classification. Through the 2003 simulation study on the face database of the Chinese Academy of Science Institute of Technology, the findings indicate that the design can achieve a higher-efficiency facial recognition and the maximum accuracy rating of 99%.

III. Proposed System

This section provides the framework of the proposed system with its components firstly. Then, it describes the most important component, which is the FR-based system from the Artificial Intelligence (AI) perspective. Finally, it presents the details of building of the FR-based system.

A. Framework of Proposed System

The framework of the proposed system consists of three main components, which as camera, the ready FR-based system, and the data base. The camera is used for face capturing, while the FR-based system takes the image of the face and processes it to be matched with the information stored in the data base. Fig. 4 shows the components of the framework.

As shown in Fig. 4, the login process will be legal if the FR-based system correctly identifies the user as an authenticated one by his face. Otherwise, the system will deny the login process. It is worth mentioning that in Fig. 4 the FR-based system is considered complete and ready for use. However, the process of building the FR-based system is described below.

B. FR-based System in Terms of AI

We can imagine that the system included in the framework described above is delivered to a company to be used by its employees. The employee is allowed to login to his or her office only if the system recognized his or her as authenticated user. In this context and in reality, the delivered system has to be built at the programmers' side and then is used at the company side. According to the fundamentals of the AI, the process of building the FR-based system goes through two main steps. The two steps are illustrated in Fig. 5.

As shown in Fig. 5, there are two main stages in the construction step, which are training and testing. In the training stage, the machine is learnt about how to recognize faces, while in the testing stage, the FR-based system is evaluated in terms of accuracy. In the usage step, a new record (face) is provided as an input to the FR-based system to test the ability of recognition (i.e., ability of dealing and handling new faces that did not train on them previously or did not stored in the data base).

C. Model Construction Step (Training Stage)

The final goal of the training stage is to train the machine to be able to recognize faces. The word "training" means that there must be a database that is used for training purpose, which in general called raw material. In this work, the raw material is represented by the database of faces. Fig. 6 shows the steps of the training stage, where the first step is to select or determine the database.

1) First step: selecting dataset: In this work, the dataset that is used for training is called ORL and obtained from [21]. The dataset contains 400 images of different faces. The images of faces belong to 40 class. The faces included in the dataset vary from persons that wear glasses to persons that has some expressions in their faces. In addition, the images of faces are taken from different angels of light. Moreover, the images are from the size 92 × 112 pixel. Fig. 7 shows the selecting data base step according to the interfaced of the system.

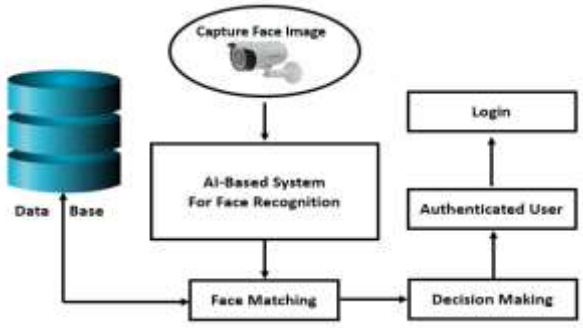


Fig. 4. Components of Framework.

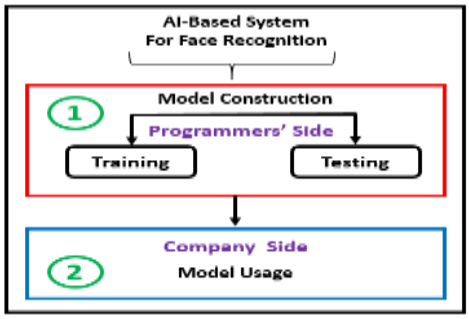


Fig. 5. Steps of Construction and usage of the System.

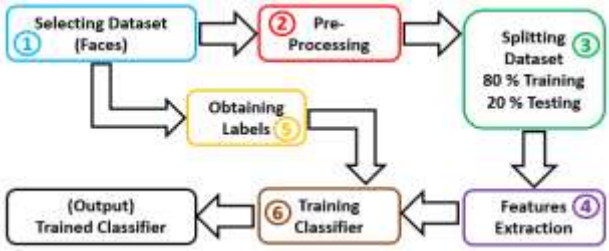


Fig. 6. Steps of Training Stage.

Fig. 7. Selecting Data base with Sample of Image.

Table I shows six images as a sample taken from the used dataset.

2) Second step: pre-processing: The objective of this step is to remove the noise from the image and to crop the face of the person, as shown in Fig. 8.

In reality, the data is noisy. Therefore, removing the noise is essential to prepare the images for training phase. In other words, the classifier will train on clean data, which in turn increases the accuracy rate. Fig. 9 shows the image shown in Fig. 7 after noise removing.

As for the technique used for noise removal, Adaptive Median Filter (AMF) is employed for this purpose. AMF contribute by adding enhancement for the mammogram input images. That is because they have the following benefits [22]:

1) Removal of salt and-pepper (impulse) noise.
2) Smoothing of other noise (may not be impulsive).
3) Reduction of distortion, such as excessive thinning or thickening of object boundaries.

Cropping process means that the face of the person located in the input image will be surrounded by a red rectangle. This in turn means that the Region of Interest (RoI) is accurately determined for further manipulation. Fig. 10 illustrates the RoI for the image used in Fig. 10.

TABLE I. SAMPLE OF IMAGES FROM DATASET

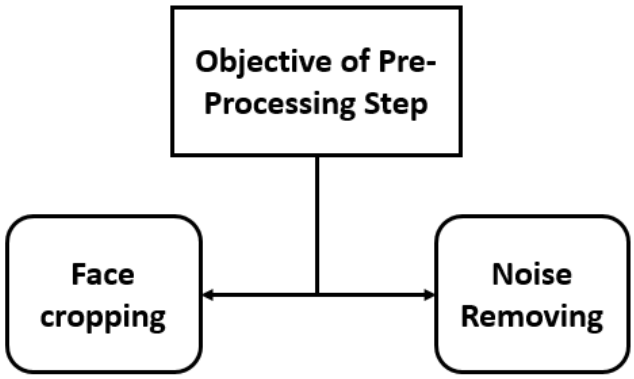


Fig. 8. Objective of Pre-Processing Step.

Fig. 9. Removing the Noise.

Fig. 10. Region of Interest (RoI).

1) Third step: splitting dataset: In this step, the original database is divided into two data sets, which are training data set and testing dataset, as shown in Fig. 11.

As shown in Fig. 11, the training dataset forms 70 % from the original data base, while the testing data set forms 30 % of the original data base. The process of splitting is performed randomly.

Fig. 11. Splitting Database.

2) Forth step: features extraction: In this work, two methods are used for feature extraction, which are Convolutional Neural Network (CNN) and Scale Invariant Feature Transform (SIFT), as shown in Fig. 12.

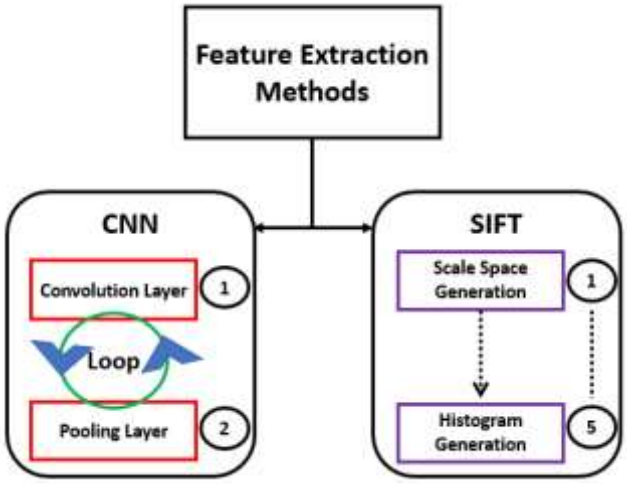


Fig. 12. Methods of Features Extraction.

As shown in Fig. 12, there are two main steps in the CNN method, while there are five steps in the SIFT method. Below in a detailed description of each method.

D. CNN based Method

The method of extracting the features depends on a loop between two main layers in the CNN, which are convolutional layers and pooling layers.

The goal of the convolution layers is to extract simple features from the input image. The goal of the pooling layers is to gather the simple features to from complete and clear features. Fig. 13 illustrates the structure of the CNN with both the convolution and pooling layers.

As shown in Fig. 13, a filter is used for feature extraction. The filter moves in a convolutional manner to scan the whole input image. The convolutional motion leads to generate the features (illustrated by the one-row tables). After features extraction, the pooling process is performed to gather the extracted features to form the final features. It is worth mentioning that the final extracted features are used for training the classifier as described in the 6th step.

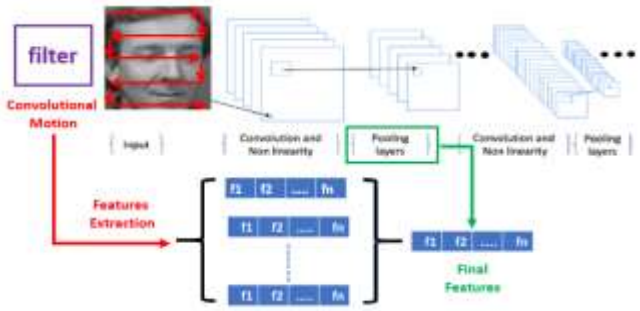


Fig. 13. Structure of CNN in Terms of Convolution and Pooling.

E. SIFT based Method

The objective of the SIFT method [23] is to extraction of features that are stationary even under change in rotation or scale of an image. In general, depending on some interesting points, the rotation invariance is guaranteed. This can be achieved by manipulation both the gradient orientations and the magnitudes of the pixels that are located as a neighbors to the interesting points. As for the scale invariance, it is guaranteed by utilizing a scale space based method. The SIFT has five steps, as illustrated in Fig. 14.

In Fig. 14, the first step is to produce the scale space. This is done by converting the face image through the Gaussian Convolution. This step aims at dealing with images as layers to be an inputs for the next step. The second step is to calculate the difference of Gaussian (DOG). This step is performed by calculating the subtracting of the nearby images. This step aims at facilitating the process of identifying the interesting points. The third step is to determine the most important interesting points. This is done by comparing neighbor pixels with the target pixels in the current and adjacent DOG images. This step aims at facilitating the process of identifying and deleting the poor interesting points (i.e., the points that have low contrast or those that are located in the edges). The forth step is calculating the gradient orientations of the neighbor pixels around the Remained Interesting Points (RIP). This step aims at determining the behavior of the interesting points in terms of directions. The final step of calculating the histogram of the RIP. The histograms are stored in vector for matching process (represented in Fig. 14 by $[a_1. a_2 \dots a_{128}]$).

Compared with the CNN based method, the SIFT based method produces less interesting points in terms of numbers. In other words, the number of interesting points obtained by the SIFT method is less than those obtained by the CNN method. This is due to the filtering process in the third step of the SIFT based method (i.e., removing poor interesting points).

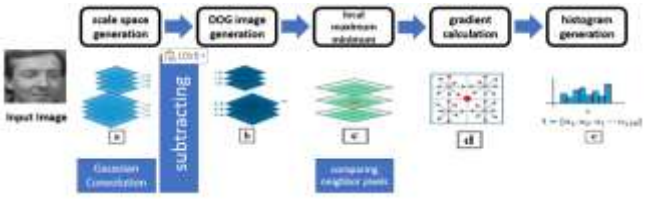


Fig. 14. Steps of SIFT based Feature Extraction Method.

Until now, the previous four steps are performed so that the process of starting training the classifier K-Nearest Neighbor (KNN) is ready. Fig. 15 illustrates that two different methods of features extraction (CNN based method and SIFT based method) are used.

3) Fifth step: obtaining labels: Before starting the process of training the KNN classifier, the labels (classes) of the used data base is obtained. The used data set has 40 class, where each class is denoted by $(S_i|i = 1.2.3 \ldots .40)$. Fig. 16 shows a snapshot of samples of classes from the used data base.

4) Sixth step: training the classifiers: In this step, one classifier is trained on the two types of features that are extracted from both the CNN and the SIFT. The classifier that is used in this work is the KNN. Since there are two methods of feature extraction, we assume that the first classifier that uses the CNN feature extraction method is called C_{CNN}^{KNN}, while the second classifier that uses the SIFT feature extraction method is called C_{SIFT}^{KNN}.

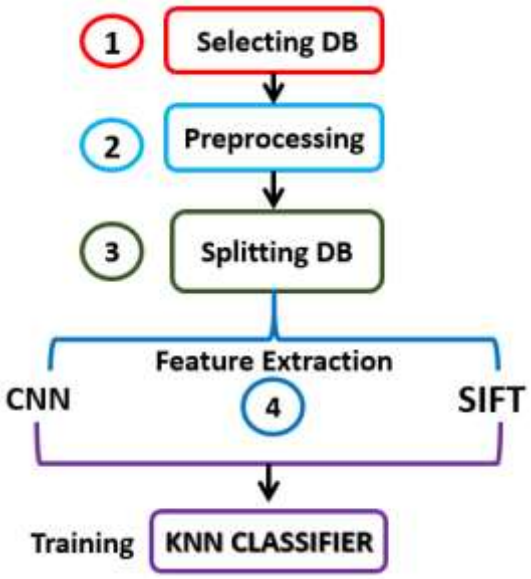


Fig. 15. Flow Chart of the Four Performed Steps.

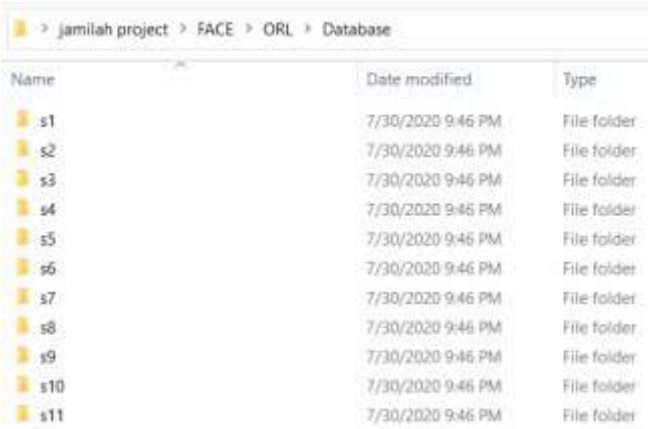


Fig. 16. Samples of Classes from the used Data Base.

F. Training the C_{CNN}^{KNN} Classifier

To train the KNN classifier, an activation function is needed. The activation function that is used in this work is the Softmax function. The reason why the Softmax activation function is used is that it has the following advantages [24]:

1) Able to handle multiple classes only one class in other activation functions normalizes the outputs for each class between 0 and 1, and divides by their sum, giving the probability of the input value being in a specific class.

2) Useful for output neurons, where typically Softmax is used only for the output layer, for neural networks that need to classify inputs into multiple categories.

Visually, the Softmax function is illustrated by Fig. 17.

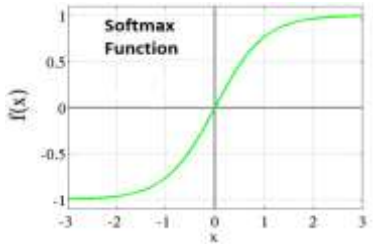


Fig. 17. Softmax Function.

As shown in Fig. 17, the Softmax function is able to represent classes within the range [-1, +1]. Since it has multiple classes property, the 40 classes found in the used data base can be represented. Fig. 18 shows the representation of the 40 classes.

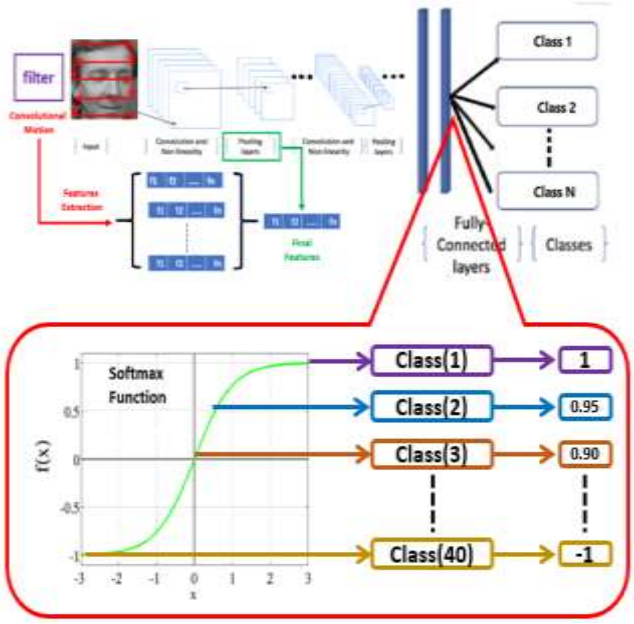


Fig. 18. Using Softmax Function for Classification..

As shown in Fig. 18, the fully connected network is formed (i.e., using the whole features extracted by the CNN). Then, the Softmax function represents the 40 class. In details, since the value that is generated by the Softmax function is limited between the -1 value and the +1 value, there is 2

degrees. To represent the 40 classes numerically, the 2 degrees is divided by 40 to calculate the step of increasing (decreasing).

$$step\ of\ increasing = \frac{2}{40} = 0.05 \quad (1)$$

Depending on the step of increasing, the first class is numerically represented by (+1) value. The 2nd class is numerically represented by (+0.95) value. The third class is numerically represented by (+0.90) value, and so on until the 40th class which represented by the (-1) value.

The KNN classifier works depending on calculating the distance between the features of given image and the center of each cluster. In other words, there will be 40 clusters. Each cluster has a center, which is represented by the value of activation function. For a given face image, the features are extracted, the value of activation function is calculated, and the distance between the value of activation function and each cluster center is determined, and finally the image is assigned to the nearest cluster. Fig. 19 shows an example for the KNN classifier.

Fig. 19 shows the CNN-KNN classifier, where the value of activation function is (-0.98). The centers of the clusters are (+1) for cluster 1, (+0.95) for cluster 2, and so on. The value (-0.98) is closer to the last cluster (its center is -1). Therefore, the image is assigned to the cluster 40.

G. Training the C_{SIFT}^{KNN} Classifier

The process of training the C_{SIFT}^{KNN} classifier is similar to the process of training the C_{CNN}^{KNN} classifier. The difference is related to using histograms as centres of clusters. Consequently, the calculating of distances is conducted between (the histogram of the face image that is under classification) and (the histograms of clusters' centres). Fig. 20 illustrates the process of training the C_{SIFT}^{KNN} classifier.

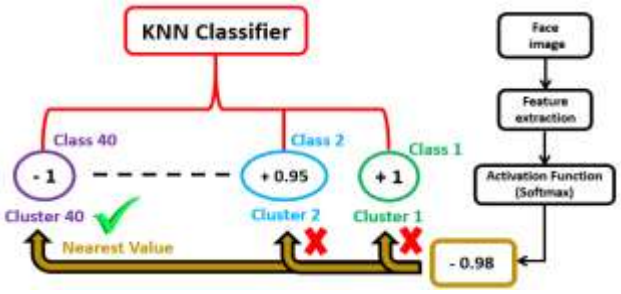


Fig. 19. CNN-KNN Classifier.

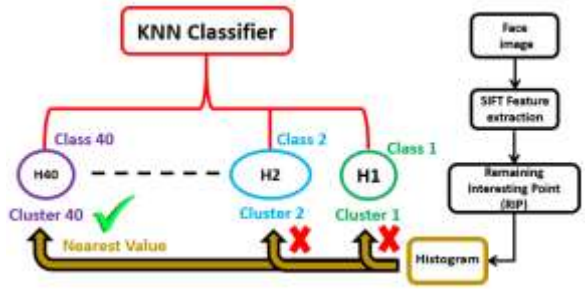


Fig. 20. SIFT-KNN Classifier.

IV. Metrics for Evaluation

To evaluate (test) the built two classifiers, the testing set is used. The built classifiers either classify a given face image correctly or incorrectly. The testing set contains 120 face images, as shown in Fig. 21.

Fig. 21. Testing the Classifiers.

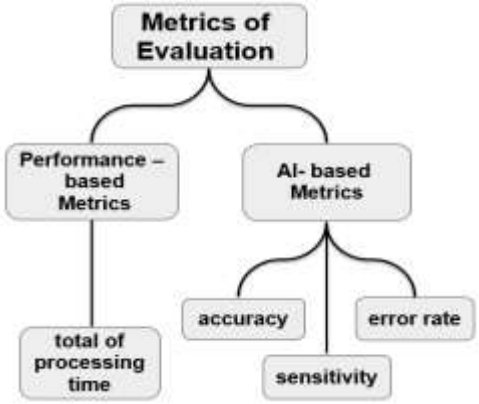


Fig. 22. Evaluation Metrics.

Two kinds of metrics are employed in the process of evaluation the proposed model. They are the AI-based metrics and the performance based metrics, as shown in Fig. 22.

In general, a confusion matrix is an effective benchmark for analyzing how well a classifier can recognize face images of different classes [25]. The confusion matrix is formed based on the following terms:

1) True positives (TP): Positive records that are correctly labelled by the classifier.

2) True negatives (TN): Negative records that are correctly labelled by the classifier.

3) False positives (FP): Negative records that are incorrectly labelled positive.

4) False negatives (FN): Positive records that are mislabeled negative.

Table II shows the confusion matrix in terms of the TP, FN, FP, and TN values.

TABLE II. Confusion Matrix

Actual class (Predicted class)	**Confusion matrix**		
	C1	**¬ C1**	**Total**
C1	True positives (TP)	False negatives (FN)	TP + FN = P
¬ C1	False positives (FP)	True negatives (TN)	FP + TN = N

Relying on the confusion matrix, the accuracy, sensitivity, and error rate metrics are derived. For a given classifier, the accuracy can be calculated by considering the recognition rate, which is the percentage of face images in the test set that are correctly classified. The accuracy is defined as:

$$Accuracy = \frac{(TP+TN)}{number\ of\ all\ images\ in\ the\ testing\ set\ (120)} \quad (2)$$

Mechanisms for accuracy-based evaluation. In this context, a higher accuracy corresponds to a better classifier output. The maximum value of the accuracy metric is 1 (or 100%), which is achieved when the classifier classifies the all face images correctly without any errors in the classification process.

Sensitivity refers to the true positive recognition rate. It is given by:

$$Sensitivity = \frac{TP}{P} \quad (3)$$

Mechanisms for sensitivity-based evaluation. In this context, a higher sensitivity corresponds to a better classifier output. The maximum value of the sensitivity metric is 1 (or 100%), which is achieved when the proportion of true positive cases equals the number of actual positive cases.

The error rate is defined as the ratio of mistakes made by the classifier during the prediction process. It is defined as:

$$eror\ rate = 1 - accuracy \quad (4)$$

Mechanisms for error rate-based evaluation. In this context, a lower error rate corresponds to a better classifier output. The minimum value of the error rate metric is 0, which is achieved when the classifier classifies all the records correctly (i.e., the accuracy is 100%).

Time dominates the situation when it comes to talking about performance metrics. In other words, the total time of stages (T_{stages}^{time}) required to build the classifier is used as a benchmark. The T_{stages}^{time} is given by:

$$T_{stages}^{time} = T_{stage}^{prep} + T_{stage}^{splittingDB} + T_{stage}^{FeEx} + T_{stage}^{OL} + T_{stage}^{Trn} \quad (5)$$

where T_{stage}^{prep} refers to the preprocessing time, $T_{stage}^{splittingDB}$ refers to the database splitting time, T_{stage}^{FeEx} refers to the features extraction time, T_{stage}^{OL} refers to the labels obtaining time, and T_{stage}^{Trn} refers to the training time. It is well known that the shorter the total time is, the higher level of performance.

V. Results and Discussions

This section is organized so that the setup is firstly presented, which describes the environment where the experiments are conducted. Then, the results with corresponding discussions are provided.

A. Setup

The context within which the experiments are conducted is shown in Fig. 23.

As shown in Fig. 23, the Matlab programming language (version: 2018) is used for implementing the proposed face recognition system. After finishing implementation stage, the proposed system is executed on a machine that has (capacity of RAM: 16 GB, and speed of Processor: 2.59 GHz). Both the CNN-based KNN and the SIFT-based KNN classifiers are involved in the comparison.

B. Results

The results are provided in practical style firstly, and then in a numerical style for more analyzing and discussion.

1) Practical style of results: The results are provided through an execution of the program, showing the results of classification in both the CNN-based KNN classifier and the SIFT-based KNN classifier. Fig. 24 and 25 shows the practical results.

The CNN-based KNN classifier shows better values in terms of accuracy (95.95 %) when compared to the SIFT-based KNN classifier (94.60 %).

2) Numerical style of results: The results are provided through the values of the AI-based metrics and the performance based metrics.

To obtain the values of the AI-based metrics, it is required to execute the program many several times. Since the testing data set contains (130 face images), it is required to execute the program 130 times using the CNN-based KNN classifier, and the same executions (on the same 130 face images) are then repeated using the SIFT-based KNN classifier. Fig. 26 illustrates the mechanism of obtaining the accuracy results.

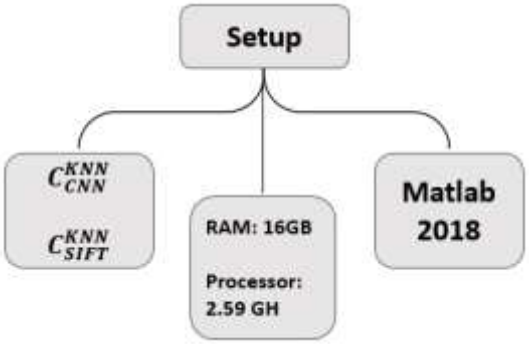


Fig. 23. Setup of Experiments' Environment.

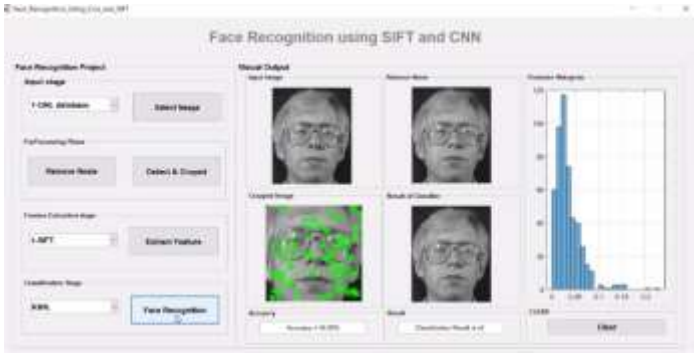


Fig. 24. Prediction of Face Class using the SIFT-based KNN Classifier.

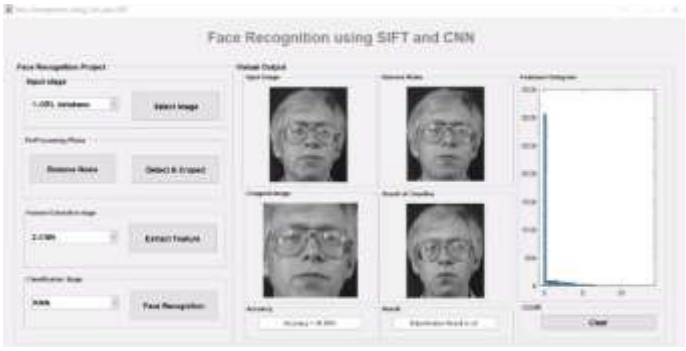


Fig. 25. Prediction of Face Class using the CNN-based KNN Classifier.

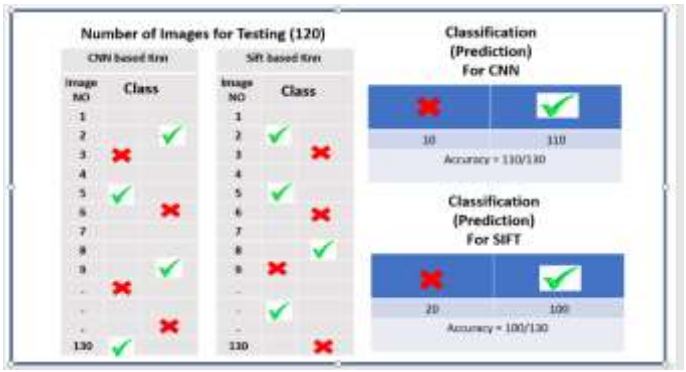


Fig. 26. Mechanism of Obtaining the Accuracy Results.

Table III shows the values of the AI-based metrics after executing the face regression system (270 time) according to the mechanism illustrated by Fig. 26.

Discussion. According to the results arranged in Table III, the CNN-based KNN classifier classified 116 face images correctly, while it misclassified 4 face images. The SIFT-based KNN classifier classified 114 face images correctly, while it misclassified 6 face images. The reason behind these results is related to the concept of both the CNN and the SIFT. In other words, the number of features that are extracted by the CNN is more than the Features that are extracted by the SIFT since the latter has a filtering phase. In addition, the CNN scans the whole face image, while the Sift scans only the space of scale that is determined. The values of the error rates support the results obtained under the accuracy metric. As for the sensitivity, the CNN-based classifier shows better values when compared to the SIFT-based classifier. That is because the sensitivity is related to how many images that are classified as true, and this increases with the increasing of the accuracy.

Fig. 27 shows the results depending on the performance-based metrics.

TABLE III. COMPARISON OF CLASSIFIERS

Classifier	**Metrics**	
	Accuracy	***Error rate***
CNN-based	97 %	3 %
SIFT-based	95 %	5 %

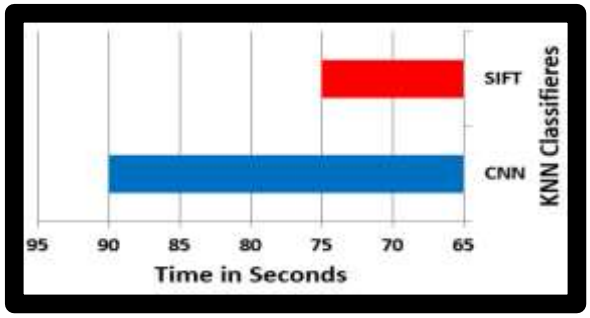


Fig. 27. Performance of the Two Classifiers.

Discussion. As shown in Fig. 27, the performance of the CNN-based KNN classifier is less than the performance of the SIFT-based KNN classifier. That is because the time required to execute the SIFT feature extraction method is shorter than the time required to execute the CNN feature extraction method. The reason behind this is related to the process of determining specific points of interest in the SIFT method. In the CNN method, the whole face image is scanned and the whole features are extracted and not be limited in specific points of interest. However, although the SIFT outperformances the CNN, the CNN achieves higher level of accuracy. In terms of cyber security, the accuracy level is preferred since the system must has the highest ability to recognize authorized users.

VI. CONCLUSION

Ensuring high level of authentication is a critical issue in cyber security systems. AI can be used to build strong face recognition systems that have the ability of providing high level of authentication as a required cyber security requirement. However, the system is poor if the degree of the accuracy is low. In this work, the CNN-based KNN and the SIFT-based KNN classifiers are proposed for face recognition. The CNN-based classifier uses the CNN itself to extract the features without using the last layers that are responsible for classification. The SIFT-based classifier uses the standard five steps of the SIFT method for feature extraction. A standard data base (ORL) is used for training and testing the classifiers. The two proposed classifiers are compared according to the accuracy, sensitivity, error rate, and time of response. The CNN-based classifier showed better results according to the AI-based metrics.

Limitation: This work did not take into consideration the performance of the CNN-based classifier. It is considered put of scope in this work since it is related to achieving high level of security.

Future work: In future work, we intend to enhance this work in terms of performance by employing Hadoop platform. In addition, the privacy and other security requirements will be taken into account.

REFERENCES

[1] Baker, Nathan, et al.Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. USDOE Office of Science (SC), Washington, DC (United States), 2019.

[2] Nixon, Mark, and Alberto Aguado. Feature extraction and image processing for computer vision. Academic press, 2019.

[3] Aileni, Raluca Maria, et al. "Cybersecurity Technologies for the Internet of Medical Wearable Devices (IoMWD)." Advances in Cyber Security Analytics and Decision Systems. Springer, Cham, 2020. 117-140.

[4] Sajjad, Muhammad, et al. "Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities." Future Generation Computer Systems 108 (2020): 995-1007.

[5] Kumar, Priyan Malarvizhi, et al. "Intelligent face recognition and navigation system using neural learning for smart security in Internet of Things." Cluster Computing 22.4 (2019): 7733-7744.

[6] Revate, S. S. "Is Biometric Security System Safe from Viral Infection Covid19?." Purakala with ISSN 0971-2143 is an UGC CARE Journal 31.9 (2020): 181-189.

[7] Basin, David, et al. "A formal analysis of 5G authentication." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

[8] Kant, Krishna. "Data center evolution: A tutorial on state of the art, issues, and challenges." Computer Networks 53.17 (2009): 2939-2965.

[9] Kemelmacher-Shlizerman, Ira, et al. "The megaface benchmark: 1 million faces for recognition at scale." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Abdullah, A. S., Abed, M. A., & Al Barazanchi, I. (2019). Improving face recognition by elman neural network using curvelet transform and HSI color space. Periodicals of Engineering and Natural Sciences, 7(2), 430-437.

[11] Khan, M. Z., Harous, S., Hassan, S. U., Khan, M. U. G., Iqbal, R., & Mumtaz, S. (2019). Deep unified model for face recognition based on convolution neural network and edge computing. IEEE Access, 7, 72622-72633.

[12] Rejeesh, M. R. (2019). Interest point based face recognition using adaptive neuro fuzzy inference system. Multimedia Tools and Applications, 78(16), 22691-22710.

[13] Gupta, S., Thakur, K., & Kumar, M. (2020). 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. The Visual Computer, 1-10.

[14] M. R. Faraji and X. Qi, "Face recognition under varying illumination with logarithmic fractal analysis," IEEE Signal Process. Lett., vol. 21, no. 12, pp. 1457–1461, 2014.

[15] B. H. Shekar and D. S. Rajesh, "Affine Normalized Krawtchouk Moments Based Face Recognition," Procedia Comput. Sci., vol. 58, pp. 66–75, 2015.

[16] P. Zhang, X. Ben, W. Jiang, R. Yan, and Y. Zhang, "Coupled marginal discriminant mappings for low-resolution face recognition," Optik (Stuttg)., vol. 126, no. 23, pp. 4352–4357, 2015.

[17] S. P. Ramalingam and P. V. S. S. R. Chandra Mouli, "Robustness of DR-LDP over PCANet for face analysis," Int. J. Multimed. Inf. Retr., vol. 7, no. 2, pp. 129–137, 2018.

[18] A. Mandhare and S. Kadam, Performance Analysis of Trust-Based. Springer Singapore, 2019.

[19] P. C. Okoye and E. A. Adenagbe, "Development of a Face Recognition System with Deep Learning and Pytorch," Int. Res. J. Eng. Technol., vol. 6, no. June, pp. 3439–3441, 2019.

[20] J. Shen et al., "Nighttime driving safety improvement via image enhancement for driver face detection," IEEE Access, vol. 6, no. c, pp. 45625–45634, 2018.

[21] kaggle(websit,2020.[line]available; https://www.kaggle.com/kasikrit/att-database-of-faces access 12septomper 2020.

[22] Saleem, S. Abdul, and T. Abdul Razak. "An effective noise adaptive median filter for removing high density impulse noises in color images." International Journal of Electrical and Computer Engineering 6.2 (2016): 611.

[23] Alberry, H. A., Hegazy, A. A., & Salama, G. I. (2018). A fast SIFT based method for copy move forgery detection. Future Computing and Informatics Journal, 3(2), 159-165.

[24] Mona Alfifi, Mohamad Shady Alrahhal, Samir Bataineh and Mohammad Mezher, "Enhanced Artificial Intelligence System for Diagnosing and Predicting Breast Cancer using Deep Learning" International Journal of Advanced Computer Science and Applications(IJACSA), 11(7), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110763.

[25] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216-231.

Regression Test Case Prioritization: A Systematic Literature Review

Ali Samad[1], Hairulnizam Mahdin[2], Rafaqut Kazmi[3], Rosziati Ibrahim[4]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, 86400 Batu Pahat, Johor, Malaysia[1, 2, 4]
Faculty of Computing, The Islamia University of Bahawalpur, 63100 Bahawalpur, Pakistan[1, 3]

***Abstract*—The techniques associated with the Test Case Prioritization (TCP) are used to reduce the cost of regression testing to achieve the objectives that the modifications in the target code would not impact the functionality of updated software. The effectiveness of the TCP is measured based on the cost, the code coverage, and fault detection ability. The regression testing techniques proposed so far are focusing on one or two effectiveness parameters. In this paper, we presented a state-of-art review of the approaches used in regression testing in detail. The second objective is to combine these effective adequacy measures into a single or multi-objective TCP task. This systematic literature review is conducted to identify the state-of-the-art research in regression TCP from 2007 to 2020. The research identifies fifty-two (52) relevant studies that were focusing on these three selection parameters to justify their findings. The results reveal that there were six families of regression TCP in which meta-heuristic regression TCP were reported in 38% and generic regression TCP techniques in 31%. The parameters used as prioritization criteria were cost, code coverage, and fault detection ability. The code coverage is reported by 38%, cost in 17%, and cost and code coverage in 31%. There were three sources for datasets were identified named Software artefact Infrastructure Repository (SIR), Apache Software Foundation, and Git Hub. The measurement and metrics used to validate the effectiveness are inclusiveness, precision, recall, and retest-all.**

Keywords—*Software testing; regression testing; test case prioritization; cost; code coverage; fault detection ability*

I. Introduction

Regression Testing (RT) is an iterative fragment of the software testing and also the primary activity during the maintenance phase. In the literature, it is mentioned that 70% of the testing cost is consumed by regression testing [1]. Once a software system is reorganized, code is modified. Whenever a software needs to be re-tested, the tester may prioritize, select or reduce the test suite size, to achieve multiple objectives of testing like code coverage, fault detection rate, cost of testing, or time. The objective of regression testing is to provide the confidence that changes did not affect the new product and reduce the overall cost of the testing. All these objectives are difficult to achieve in a single testing cycle. If the coverage should increase, cost and time also increased [2]. In regression testing, 100% of code coverage may not be preferred. The efficiency of prioritization may raise the yield of the testing procedure by the means of fault detection ability.

RT helps in testing the code by analyzing the target code both in original and updated form. Furthermore, it performs checking with the assumption that the updates in the target code has minimum or negligible effects on the services provides by the software [3]. The reports claim that code testing is 80% of the total cost of the software cost which is different from the maintenance cost that is about 50% of the total cost [4-6]. One of the objectives of regression testing is to reduce the testing cost by using the state-of-art approaches used in Test Case Selection (RTS), Test Case Prioritization (TCP) along Test Case Reduction (RTR) [7].

The classic techniques of TCP consist of three general components, TCP framework, prioritization parameters and prioritization adequacy measures as shown in Fig. 1. This generalized process takes original program P, modified program P′ and test suites T as input. The prioritization process may have a framework which identifies the code change information from P and P′ and other relevant information like code coverage, fault detection ability, and executional cost. The test case prioritization measure may prioritize the test cases from T and move them to T′ (a subset of T), based on computations performed by selection logic described in the framework. The TCP adequacy measures are used to assess the effectiveness of TCP technique and results produced by this technique.

The TCP adequacy measures are effective in judging the effectiveness of the TCP process. These measures are computed in two ways in the TCP process, the first is to prioritize the test cases on these prioritization contexts like TCP based on coverage measures, TCP based on cost measures, and TCP based on fault rates. The second use is to assess the effectiveness of TCP by coverage or cost optimization. There are four challenges to organizing the three parameters (cost, coverage, and fault detection) in a fashion to assess their importance, dependencies, and priority concerning each other. The other challenge is to choose the appropriate type of these measures, like coverage subtypes, cost subtypes, fault types, and severity. The third challenge is to identify the relevant frameworks that adjust the three parameters for the prioritization of test cases and their adequacy scale and use as adequacy measure. The fourth challenge is to identify the techniques based on these effectiveness measures, such as execution cost, code coverage, and fault detection ability. The primary objective of this paper is to define the TCP effectiveness based on cost, code coverage, and fault detection ability as effectiveness contributors. Furthermore, this survey assesses the current state-of-the-art algorithms in the design of

the regression test case prioritization frameworks and techniques so far. The secondary objectives of this research are to identify the available datasets and methods for the solution of test case prioritization problems.

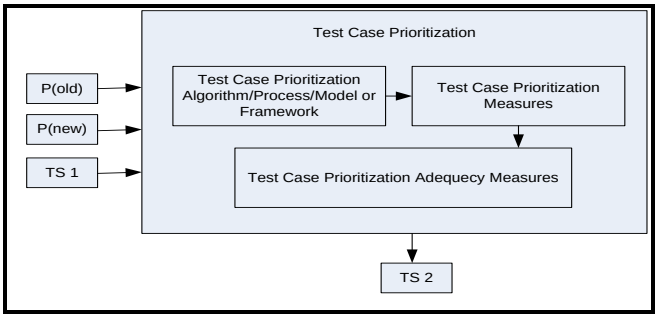


Fig. 1. Maintenance Process Model.

The rest of the paper is organized as follows: Section II formulates the literature selection process of the studies; Section III encompasses data extraction and Section IV includes the related work that reviews operational profiles. Finally, Section V concludes this research.

II. Systematic Literature Review Process

To conduct this SLR, three guidelines are followed [1-3]. These guidelines provide the steps to conduct the literature review. The SLR method of conducting a literature review is borrowed from clinical research to organize the data from previous research and systematically deducing the results. The sub-sections include these step by step details to conduct the research process. These steps are review protocol, framing research questions, the primary studies selection, search keyword selection, inclusion and exclusion criteria for primary studies and results and synthesis based on selected primary studies. Initially few papers were handpicked seeing the titles and abstracts. Then, a citation based on forwarding snowballing strategy was adopted [16], computing inclusion and exclusion criteria and examining search statistics of the focused domain. In the subsequent stage, specialized search queries were formed to gather the studies that satisfied the inclusion-exclusion criteria and their match relevance.

A. Review Protocol

The SLR review protocol helps us to execute this research process with necessary actions and outputs. The SLR research protocol is shown in Fig. 2. The SLR process is started to provide the rationale for the purpose and need of study. The research questions are framed to collect the data for the purpose of fulfilling the research objectives. The next step is to collect the primary studies, inclusion and exclusion criteria, which helps to collect the most relevant research studies with respect to the research questions framed for this SLR. The data extraction method is devised to collect data from primary studies and then finally the data has been collected for synthesis and analysis purpose.

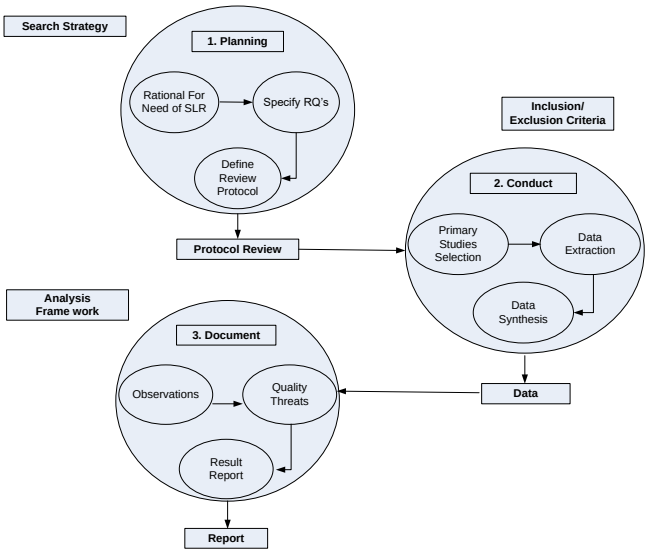


Fig. 2. The Review Protocol for Systematic Literature Review.

B. Research Questions

The research questions are framed with the help of discussions with the domain experts and software testing literature blind searches. The primary focus of these research questions is to find out the most relevant research on regression based test case prioritization adequacy criterions for test case prioritization, datasets available and used in controlled experiments for test case prioritization, measurements, and metrics available for regression testing and test case prioritization. The focus of these research questions was also to find those test case prioritization techniques which use more than one or two prioritization parameters and the effects of these parameters on the results of these techniques. The SLR also tries to focus on effectiveness as a measurable fact which so far discussed in the literature as a qualitative fact instead of a quantitative parameter [3, 12, 13]. The research questions are shown in Table I, with their justification to include in this SLR.

C. The Study Selection Procedure

The most important part of an SLR is its selection of primary studies which provides the ground for synthesis and analysis of data. The objective is to collect the most relevant data for results that identify the domain trends and dominant research problems with their solution space [3]. The quality of results based on the relevance of these primary studies. The selection of primary studies for this SLR based on the following steps.

- The selection of research repositories.
- The formulation and choice of keywords for search queries.
- The inclusion and exclusion criteria for searched studies with respect to the research questions.

TABLE I. THE RESEARCH QUESTIONS FOR SYSTEMATIC LITERATURE REVIEW

No	Research Questions	Justification
RQ-1	What is the state of the art research in regression TCP types/techniques?	The objective of this research question is to identify the important trends in the regression TCP research domain in order to collect the evidence for design and analysis for new and emerging regression TCP techniques.
RQ-2	What are the selection parameters used in regression TCP techniques?	The objective of this research question is to identify all possible selection parameters for regression TCP techniques and to find why they are used as selection criteria for test case selection.
RQ-3	What type of datasets used in regression TCP experiments?	The purpose of this research question is to find out the datasets for regression TCP experimentation and their usage.
RQ-4	What type of metrics/ evaluation criterions are used to verify the regression TCP techniques?	This research question helps to identify the possible metrics to evaluate and verify the regression TCP techniques and methods.

1) The selection research repositories: The process to identify the primary studies has been initiated by randomly entering the search keywords to research repositories. These retrieved research studies are then compared to the objectives of the research questions and inclusion/ exclusion criterion has been applied to these retrieved research studies. The choice of research repositories is quite important because of the quality dependent on these choices. For this purpose, in mind, the authors used the following research repositories is used for this process.

a) Science Direct.

b) IEEE Explore.

c) ACM Library.

The choice of these repositories based on the fact that IEEE Explore and ACM Library contains almost every important conference in the software testing domain. The Science Direct contains the research studies of almost all important journals relevant to the software testing research domain [4, 5].

2) Search keywords selection: A precise and systematic approach has been devised to search the search keywords. The approach is comprising of the following steps.

a) The most repeated keywords are selected from review papers on software testing and regression testing.

b) Find out the matching words, alternative keywords, similar words for these most frequently used terms in software testing literature.

c) Then devised search strings and search queries by using AND, OR and NOT operators available in research repositories search engines.

d) In the last step, we apply manual verification on searched studies that the research studies are relevant to the research questions.

In order to collect the most relevant research studies, authors try to switch the keywords with OR operator with author titles and author keywords are switched. The time period is also defined from 2007 to 2019 to limit the number of studies and covering the last twelve years of progress in the domain. This time limit was applied to the reason that software testing has a tremendous amount of research papers, but the systematic methodology was adopted in the year 2007, so it is helpful to limit the most relevant studies by applying this time limit. The search queries are shown in Table II.

TABLE II. THE SEARCH QUERY FOR SYSTEMATIC LITERATURE REVIEW

Repository	Search Query
IEEE	((((((("Publication Title":test case prioritization) OR "Abstract":test case prioritization) OR "Author Keywords":test case prioritization) OR "Publication Title":test suite prioritization) OR "Author Keywords":test suite prioritization) OR "Abstract":test suite prioritization) Filters Applied: Conferences Journals 2007 - 2019
ACM	"query": { acmdlTitle:(+Test +case + prioritization) OR recordAbstract:(+Test +case + prioritization) OR keywords.author.keyword:(+Test +case + prioritization) "filter": {"publicationYear":{ "gte":2007 }}, {owners.owner=HOSTED}
Web of science	TITLE: (Test case prioritization) OR TITLE: (Test suite prioritization) Timespan: years 2007. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI.

3) Inclusion and exclusion criteria for searched studies: The regression test case prioritization has many different objectives with application domains, testing scope and testing environments. The test case prioritization in general considered as test suite optimization technique, but it is also observed that optimization research has many viewpoints, applications other than software testing. The challenge in study selection was the diversity of the topics covered under software testing such as software test suite prioritization, reduction, and augmentation. The experimental scope and size is also the main concern while selecting primary studies. As it was stated that the focus of this SLR was to collect the evidence for research studies considering cost, coverage and fault detection ability as test case prioritization criteria and effectiveness of the proposed techniques must be considered as one of the objectives of these studies. Therefore, it was required to design some rules while including or excluding the searched studies. Here, the inclusion and exclusion criteria applied to search studies were discussed.

a) The search queries are applied to selected research repositories and found 855 research studies. Then the authors applied two-stage inclusion criteria as shown in Table II.

b) The studies must be in the English language.

c) On the first level, the studies selected which have test case prioritization, test suite optimization with test case prioritization, test suite effectiveness, cost/coverage/fault

detection based test case prioritization in their title are selected.

d) The studies not included test case prioritization, test suite effectiveness or test case prioritization with some optimization criteria in their title or abstract are excluded.

e) The research studies that are not experiments, controlled experiments, case studies or without empirical results are also excluded.

Table III presents a two-Stage Spectrum of Research Studies Inclusion/Exclusion.

TABLE III. THE TWO STAGE SPECTRUM OF RESEARCH STUDIES INCLUSION/EXCLUSION

Research DB	First Searched Studies	First Round Exclusion	Second Round Inclusion
Science Direct	135	23	15
ACM Library	623	108	8
IEEE Explorer	900	260	29
Total	1658	391	52

After the first level of exclusion/inclusion, the authors started the second level of exclusion/inclusion. In this phase the studies are organized as per the research question framed. The content of each research study is compared with the objectives of the research questions especially the experimental process and result section of the study. The studies are now excluded/included based on the following rules.

1) The studies that did not report any experimental, case study or controlled experimental results are excluded.

2) The studies less than five pages and without experimental details are excluded.

3) The posters, PhD or Master thesis are excluded.

4) The technical reports are excluded.

5) The studies that did not focus on test case prioritization, test case prioritization optimization is excluded.

6) The studies that did not consider cost, coverage and fault detection ability as prioritization criteria or effectiveness criteria are excluded.

The purpose of the second phase of exclusion was to collect the most relevant and reasonably high-quality research studies with some experimental insights towards the domain. After second phase of inclusion/exclusion, authors left with fifty-two research studies that focus on regression test case prioritization with focus on cost, coverage or fault detection ability as test case prioritization parameters or used as effectiveness measure from these three parameters (cost, coverage and fault detection ability).

D. The Data Collection Strategy

After collecting the most relevant studies from inclusion/exclusion criteria, the data collected from these studies have been followed [6, 7]. The data also collected into two phases. The first phase consists of a study title, publication year and source, summary of the research study and comments of the researcher. In the second phase, the technical information with respect to the research questions has been collected to answer the research questions framed for this SLR. The first phase data collection helps the researchers to execute the inclusion and exclusion phase. The second phase of data collection helps the researchers to synthesis and analysis the results of this SLR.

III. RESULTS AND DISCUSSION

In this section, the results are presented based on the data collected from the primary studies to answer the research questions framed in the previous section. There were four questions framed for synthesis and analysis. These questions were framed to identify the main research gaps and important trends in the development and design of the regression test case prioritization research domain. The second focus was to identify the datasets and experimental evaluation trends and features in the regression test case prioritization research domain. The results for Research Question 1: The state of the art research in each research questions are as followed regression test case prioritization types/techniques.

The objective of this research question was to assess the state-of-the-art research conducted in the domain of regression test case prioritization with the focus on cost, coverage and fault detection ability. The analysis performed on the data collected from primary studies shown the following regression test case prioritization techniques families as in Fig. 3. Each technique has common input, processing, and output styles but differs in their designing parameters and context of usage.

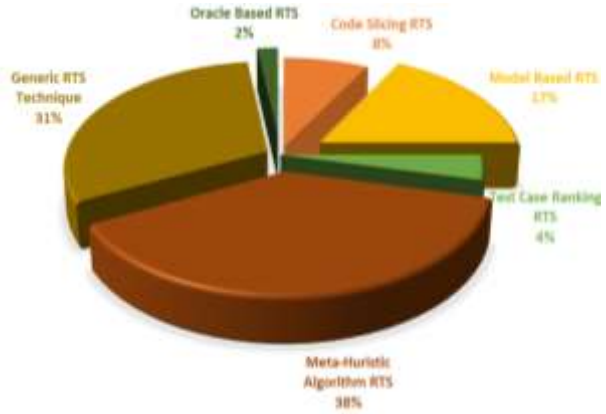


Fig. 3. The RTS Techniques Classification based on Primary Studies.

From Fig. 3, the major families of regression test case selection techniques as follows.

1) The meta-heuristic based TCP
2) Model-based TCP
3) Generic Based TCP techniques.
4) The test case ranking based TCP.
5) The Code slicing based TCP.
6) The Oracle Based TCP.

The meta-heuristic based TCP techniques were found 38%, as a leading trend in TCP methods. There were 20 out of 52 studies that used these algorithms to solve or implement the solution for the TCP problems. The reason for its popularity was its capability to handle the multi-criteria problems with emerging tools and technologies for analysis and design for

these algorithms. From these meta-heuristic family Genetic Algorithm (GA) was observed in seven (7) out of twenty (20) studies and become the most widely used algorithm for TCP problems. The GA is considered as evolutionary optimization technique with the inbuilt believe in survival to the fittest. The GA is popular for TCP solution design due to its nature of selection the stronger population based on some fitness function. The design and process of GA very much like TCP design and process of selection. TCP selects the test cases to prioritize from already used test suites based on some criteria while GA selects the stronger population from previous populations based on fitness functions. The second reason for the choice of GA for TCP problems was its maturity and there were so many comparative studies available for this algorithm. There are many datasets available with evaluation metrics with GA in the test case prioritization research domain. The different types of GA used in these studies are the Co-evolutionary Genetic Algorithm (CGA), Diversity Based Genetic Algorithm (Div-GA), Multi-Objective Genetic Algorithm (MOGA) and Non-Dominated Sorting Genetic Algorithm (NSGA-II).

The second most used algorithm in TCP problem solving was Particle Swarm Optimization (PSO) observed in five (5) studies out of twenty (20) studies. PSO is a greedy algorithm that tries to find a local maximum from the problem space. It is easy to implement as compared to GA. But the choice between GA and PSO depends on the nature and design of the problem. The different types of implementations of PSO from primary studies are simple PSO, Multi-objective PSO and Additional Greedy based on voting mechanism PSO.

The fuzzy algorithm is the third most used algorithm four (4) out of twenty studies. The fuzzy is used with types of rule-based fuzzy, fuzzy classification and fuzzy expert system. The fuzzy is quite a simple but static decision-making system. The prior defined rules are used to decide the different decisions required during the selection of TCP. The K-means and semi-supervised clustering also used in two different studies for TCP problems.

The second class of solutions for TCP problems were Generic TCP solutions found in 31% of the studies. There were sixteen (16) out of fifty-two studies (52) studies that used these methods, tools, and algorithms. These are self-designed custom solutions for specific tools and problems. Normally they are applied to industrial-scale case studies to solve TCP problems.

Model-based TCP is the third popular class of TCP solutions. It was used in nine (9) studies out of fifty-two (52) studies. It is based on Unified Modeling Language (UML) artifacts to prioritize the test cases for software under testing. The used artifacts were activity diagrams, state machines, and use case diagrams. But these diagrams appear so early in the software life cycle, so, they are so much imprecise to use as test case prioritization solutions. The code slicing and chopping techniques are used in four (4) studies out of fifty-two (52) studies, 8% of the total studies. These techniques were relevant due to code modifications are the primary focus of regression TCP techniques. The code changes and modifications are easy to identify by code slicing and code chopping techniques. But due to the complexity of new coding environments, it is difficult to chop the code with modern code editors and code generators. The test case ranking regression TCP techniques were seen in two out of fifty-two studies. The test cases and their results were used to rank the test cases for future use in these techniques.

A. *Research Question 2 The Selection Parameters used in RTP Techniques*

This research was framed to identify the number of parameters used for test case prioritization techniques. The objective was to understand the fact that available space for research in designing new test case prioritization techniques, their design trends and the dependency among these parameters if there is any dependency among these parameters. The well-known parameters are cost, coverage and fault detection ability [8]. The definitions of these measures are as following.

1) Cost: The time or resources consumed by a test suite/ test case to complete its execution on source code to return its results. The further types of cost observed are time to run a test suite, time to create a test suite, time to analysis for a test suite and time to prepare the results of a test suite.

2) Code coverage: The ratio of source code executed by a test case/test suite to the total number of source lines expected to execute by that test suite/test case is known as code coverage. Its special sub-groups are statement coverage, condition coverage, modified condition coverage, loop coverage, branch coverage, modified branch coverage, and modified statement coverage.

3) Fault detection ability: The number of the faults identified by a test suite/test case is known as fault detection ability of that test suite/test case. The sub-types of faults observed in primary studies are structural faults, real faults, hand seeded faults and mutation faults.

The results of the research question are shown in Fig. 4 below. The observed classes of these prioritization criteria are cost, code coverage and fault detection ability as single criteria to test case prioritization techniques. The code coverage with Fault detection ability and cost and code coverage are observed as bi-criteria test case prioritization parameters. The cost, code coverage, and fault detection ability are observed as tri-criteria test case prioritization parameters.

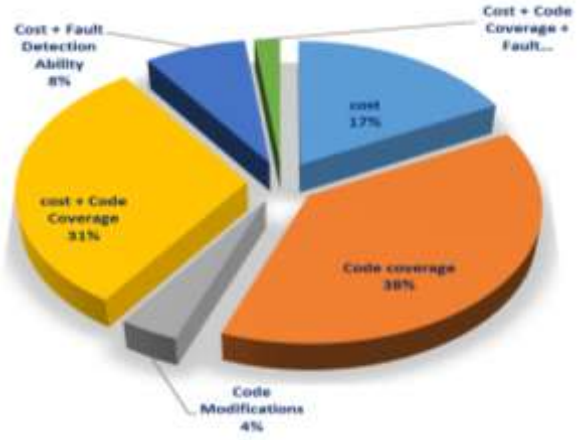


Fig. 4. Test Case Prioritization Parameters Classification based on Primary Studies.

The code coverage is the most dominant trend observed in selected primary studies. It was found in twenty (20) out of fifty-two (52) primary studies which were 38% of the total primary studies. The reasons for using code coverage were its simple computation with respect to other parameters. There were a good number of the tools available for measuring code coverage of different types and its integration is quite simple with available code editors and code generators like Eclipse, Junit, Code Cover, Mue-java, etc. The measurement of code coverage is simple enough and decision making is also very straight forward. The more code coverage provides more confidence in testing teams that their code is tested. The code coverage is used as a proxy in many test scenarios which means for testing teams, quality assurance groups, management teams and customers of the product.

The second dominant test case prioritization parameter group was cost and code coverage, which is also a bi-criteria test case prioritization family. It was observed in sixteen (16) out of fifty-two (52) primary studies, which was 31% of the total primary studies. The reason for this was in close resemblance in the measurement of these parameters. Both code coverage and cost metrics returned the results in measurable numbers. The available tool support for measuring cost and code coverage. The code coverage and cost measurement both dependent on each other, more coverage means more cost for a test suite. The more cost means a less effective regression TCP technique. Both code coverage and cost were primary objectives for the optimization of test case prioritization techniques.

The third trend found in primary studies was cost-based test case prioritization techniques. It was found in nine (9) out of fifty-two (52) primary studies which were 17% of the total primary studies. The optimization of the cost was the primary objective of a test case prioritization technique because the reduction in cost means a better test case prioritization technique which may replace the previous regression TCP technique. The cost measures are observed with many different viewpoints like execution cost of a test suite, size of the source code under testing, size of the test cases in a test suite, analysis time for results of a test suite, preparation of the test suite for a software under testing and post-analysis and prioritization time of a test suite. The choice of cost measures depends on the local requirements of optimization of test case prioritization problems. The fourth trend was the tri-criteria test case prioritization parameter comprises of three measures cost, code coverage, and fault detection ability. The combination of these three parameters makes test case prioritization more effective because fault detection ability is the primary objective of all software testing techniques. The fault detection ability in test case prioritization techniques used as adequacy criteria so far, but in a few techniques, it was used as test case prioritization parameters as well. The tri-criteria optimization seen as a challenge in test case prioritization problems due to the huge size of the code and test suite sizes for software under testing. The last test case prioritization parameter was code modifications, the identification of code changes from code in code chopping techniques. The code chopping was not practical due to increase in size and complexity of the source code in modern software. The second reason was the security and safety requirements of the third-party source codes which may not provide direct access to the critical pieces of the source codes.

B. *Research Question 3: Datasets used in Regression TCP Experiments*

The research question was framed to identify the datasets used in software testing experimentation with a special focus on designing the novel techniques for regression testing. The software testing datasets are quite different in nature as compared to other artifacts used in software engineering research. The point of differences and important features considered during datasets for regression testing are as follows continue Table IV.

1) There must be a reasonable source code size for the software under testing.

2) There is must be a test suite available for testing with previous testing cycle's history or results which justify the usage of that test suite.

3) There must be some tool/framework/methodology support available to execute that testing technique on software under testing.

4) There should be some measurement mechanism to evaluate and compare the results for that testing technique.

5) The source code and test suite collections must available to other research communities to use as an artefact for their experiments.

6) The results and conclusions must be based on some environment available to other research communities to evaluate and compare with their findings with the previous research findings.

There were three sources identified providing software source code, test suites, test results and tool information used to collect the results for software testing experiments. These sources are as follows:

1) SIR (Software Artefact Infrastructure Repository).
2) Open Source (Apache Software Foundation).
3) Git-Hub.

The Software-Artefact Infrastructure Repository (SIR) [9] is the collection of software source codes with multiple versions and associated test suites. It has the artifacts that have a wide range of software with many different programming languages like Java, C, C++, PHP and C-sharp. These datasets are prepared for unit testing, integration testing, system testing. The fault types supported by these datasets are real faults, hand seeded faults and mutation faults [9]. The detailed primary studies and subject software are listed in Table IV.

The second repository which offers a wide range of datasets for software testing artifacts is Apache Software Foundation [61]. This repository contains 200 Million lines of source code and 350 projects with multiple versions of source code and test suites for each version. The Git Hub is also a very huge size code repository for software source codes and their test suites. The third repository Git Hub [62] is a general-purpose repository in which individual developers and software engineers upload their code and test suites. The choice of datasets depends upon the nature and design of the problem,

available tool support for technique under analysis and measurement methods used to evaluate results produced with these datasets.

TABLE IV. THE DATASETS IDENTIFIED FOR RTS EXPERIMENTS IN PRIMARY STUDIES

Study	Reference	Dataset
1	[10]	Custom Product/Code
2	[11]	Custom Product/Code
3	[12]	SIR (Siena, Jtopas)
4	[13]	Not Reported
5	[14]	Not Reported
6	[15]	Custom Product/Code
7	[16]	Aspect Compiler example package = 3 programs
8	[17]	ABB = 3 programs, SIR = 2 programs
9	[18]	ABB program
10	[19]	Custom Product/Code
11	[20]	SIR (Jmeter), XML(Security, ANT)
12	[21]	Custom Product/Code
13	[22]	Custom Product/Code
14	[23]	SIR (Flex, Space, Schedule)
15	[24]	Student Enrolment System.
16	[25]	SIR (Nano,ant,Galileo, Jmeter,XML)
17	[26]	SIR (nanoXML,jtops,jmeter,xml- security, any)
18	[27]	Custom Product/Code
19	[28]	Safety Monitoring Component
20	[29]	Open-Source (Apache,Log 4j, common-Math)
21	[30]	Open-Source (Polo)
22	[31]	SIR(Space)
23	[32]	Custom Product/Code
24	[33]	Video-conference system Safety Monitoring Components
25	[34]	Not Reported
26	[35]	Microsoft Dynamics AX
27	[36]	Not Reported
28	[37]	Custom Product/Code
29	[38]	Scheduler
30	[39]	SIR(print tokens, printtokens1, scheduler, scheduler2, space)
31	[40]	Custom Product/Code
32	[41]	SIR 11 programs
33	[42]	SIR(printtokens, printtokens2)
34	[43]	Custom Product/Code
35	[44]	SOFIE is the tax accounting system
36	[45]	Custom Product/Code
37	[46]	Custom Product/Code
38	[47]	Not Reported
39	[48]	SIR (space)
40	[49]	Calendar, triangle, time-date,Kmap generation, tax calculation.
41	[50]	Custom Product/Code
42	[51]	Custom Product/Code
43	[52]	Custom Product/Code
44	[53]	11 Large Open-source projects
45	[54]	61 open source systems
46	[55]	11 open-source projects
47	[56]	21 Java projects
48	[57]	Grep v1 to v7
49	[58]	Custom datasets for 3 sprints
50	[59]	JFree Chart, Apache Tomcat, Argo UML
51	[60]	37 projects on Git Hub

C. *Research Question 4: Type of the Metrics/ Evaluation Criterions are used to Verify the Regression TCP Techniques*

The research question was framed to identify the measurements and methods to evaluate the results collected from regression TCP experiments. The important features are those which classify the effectiveness abilities of one technique to another technique. There are so many different viewpoints observed from primary studies collected for this SLR. The notable trends were comparing the results in terms of their input, process and output styles, their method to prioritize the test cases, their ability to identify the faults and fault types and their presentation method of the finding for analysis performed on datasets.

The code-based regression TCP techniques primarily designed to reduce the cost of testing in terms of execution time, test suite size and try to satisfy the code coverage required criteria and cost evaluation and fault detection ability. In the evaluation of regression TCP experimental results, a framework has been proposed in the study [63]. This evaluation structure classifies the test suites in the following types.

1) Obsolete Test Cases: A test case that uncovers nothing new like faults or code modifications.

2) Modification Revealing: A test case that executes a modified part of the code under testing.

3) Non-Modification Revealing Test Case: A test case that does not execute a modified part of the code under testing.

4) Fault Revealing Test Cases: The test cases which identify the faults from the source code under testing.

The other metrics identified from the primary studies are inclusiveness, precision, fault measure, fault rate, code coverage, fault metrics, and retest-all. The studies are compared in terms of effectiveness, cost, code coverage, and fault detection ability. These metrics are mentioned because there were proper mathematical grounds for these metrics were available. The second reason was that many experiments reported from primary studies may be useful to compare the results with future studies.

IV. Conclusion and Future Work

The research study was conducted by following a systematic literature review methodology. The review protocol was designed and conduct the search from relevant research repositories with the research questions framed in the review protocol. There were 1658 studies found from three research repositories. There were two-stage inclusion/exclusion criteria to choose the most relevant studies with respect to the research questions. There were 391 studies left on the first level of inclusion/exclusion criteria. On the second level of inclusion/exclusion criteria, there were fifty-two studies left. The analysis of primary studies reveals that there are six main classes of test case prioritization techniques such as meta-heuristic regression TCP, code slicing regression TCP, model-based regression TCP, test case ranking regression TCP, Oracle-based regression TCP, and Generic regression TCP techniques. The regression TCP parameters have cost, coverage and fault detection, as single criteria regression TCP, the cost and coverage and fault and coverage as bi-criteria regression TCP and cost, coverage and fault detection as tri-criteria regression TCP techniques. There was a long list of datasets available for controlled experiments of regression testing experiments. The main sources to obtain these datasets were SIR, Git Hub, and Open Source Apache Software Foundation. It is also concluded that meta-heuristic techniques are the most researched trend so far. The genetic algorithm was the most used algorithm for regression TCP solutions. The code coverage is the most used parameter for test case prioritization. Based on these results, the authors recommend that more experimental research is required to investigate bi-criteria and tri-criteria test case prioritization techniques. It is also concluded that cost, coverage, fault detection ability and code modifications are equally important for selecting a test suite for software under testing. The results reviewed from primary studies show that these studies ignore one or two prioritization parameters. It is also observed that the local constraints like tools, programming languages, and measurement and metrics also need to be researched experiment to produce generalized results for the whole testing community.

References

[1] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," Journal of systems and software, vol. 80, pp. 571-583, 2007.

[2] 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.

[3] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," Information and software technology, vol. 51, pp. 7-15, 2009.

[4] M. Younas, D. N. Jawawi, I. Ghani, T. Fries, and R. Kazmi, "Agile development in the cloud computing environment: A systematic review," Information and Software Technology, vol. 103, pp. 142-158, 2018.M. Khatibsyarbini, M. A. Isa, D. N. Jawawi, and R. Tumeng, "Test case prioritization approaches in regression testing: A systematic literature review," Information and Software Technology, vol. 93, pp. 74-93, 2018.

[5] B. Kitchenham, H. Al-Khilidar, M. A. Babar, M. Berry, K. Cox, J. Keung, et al., "Evaluating guidelines for reporting empirical software engineering studies," Empirical Software Engineering, vol. 13, pp. 97-121, 2008.

[6] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, et al., "Preliminary guidelines for empirical research in software engineering," IEEE Transactions on Software Engineering, vol. 28, pp. 721-734, 2002.

[7] M. Huang, S. Guo, X. Liang, and X. Jiao, "Research on regression test case selection based on improved genetic algorithm," in Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on, 2013, pp. 256-259.

[8] H. Do, S. Elbaum, and G. Rothermel, "Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact," Empirical Software Engineering, vol. 10, pp. 405-435, 2005.

[9] C. Tao, B. Li, X. Sun, and C. Zhang, "An approach to regression test selection based on hierarchical slicing technique," in Computer Software and Applications Conference Workshops (COMPSACW), 2010 IEEE 34th Annual, 2010, pp. 347-352.

[10] W.-T. Tsai, X. Zhou, R. A. Paul, Y. Chen, and X. Bai, "A coverage relationship model for test case selection and ranking for multi-version software," in High Assurance Systems Engineering Symposium, 2007. HASE'07. 10th IEEE, 2007, pp. 105-112.

[11] C. Tao, B. Li, X. Sun, and Y. Zhou, "A hierarchical model for regression test selection and cost analysis of java programs," in Software Engineering Conference (APSEC), 2010 17th Asia Pacific, 2010, pp. 290-299.

[12] W. S. A. El-hamid, S. S. El-etriby, and M. M. Hadhoud, "Regression test selection technique for multi-programming language," in 2010 The 7th International Conference on Informatics and Systems (INFOS), 2010, pp. 1-5.

[13] S. Huang, Z. J. Li, J. Zhu, Y. Xiao, and W. Wang, "A novel approach to regression test selection for J2EE applications," in 2011 27th IEEE international conference on software maintenance (ICSM), 2011, pp. 13-22.

[14] Z. Xu, Y. Liu, and K. Gao, "A novel fuzzy classification to enhance software regression testing," in Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on, 2013, pp. 53-58.

[15] G. Xu and A. Rountev, "Regression test selection for AspectJ software," in Software Engineering, 2007. ICSE 2007. 29th International Conference on, 2007, pp. 65-74.

[16] T. Yu, X. Qu, M. Acharya, and G. Rothermel, "Oracle-based regression test selection," in Software Testing, Verification and Validation (ICST), 2013 IEEE Sixth International Conference on, 2013, pp. 292-301.

[17] J. Zheng, L. Williams, B. Robinson, and K. Smiley, "Regression test selection for black-box dynamic link library components," in Proceedings of the Second International Workshop on Incorporating COTS Software into Software Systems: Tools and Techniques, 2007, p. 9.

[18] P. K. Chittimalli and M. J. Harrold, "Recomputing coverage information to assist regression testing," Software Engineering, IEEE Transactions on, vol. 35, pp. 452-469, 2009.

[19] N. Rachatasumrit and M. Kim, "An empirical investigation into the impact of refactoring on regression testing," in Software Maintenance (ICSM), 2012 28th IEEE International Conference on, 2012, pp. 357-366.

[20] A. Pasala, Y. L. Y. Fung, F. Akladios, G. A. Raju, and R. P. Gorthi, "Selection of regression test suite to validate software applications upon deployment of upgrades," in 19th Australian Conference on Software Engineering (aswec 2008), 2008, pp. 130-138.

[21] E. Fourneret, J. Cantenot, F. Bouquet, B. Legeard, and J. Botella, "Setgam: Generalized technique for regression testing based on uml/ocl models," in 2014 Eighth International Conference on Software Security and Reliability (SERE), 2014, pp. 147-156.

[22] A. Nanda, S. Mani, S. Sinha, M. J. Harrold, and A. Orso, "Regression testing in the presence of non-code changes," in 2011 Fourth IEEE International Conference on Software Testing, Verification and Validation, 2011, pp. 21-30.

[23] S. Chen, Z. Chen, Z. Zhao, B. Xu, and Y. Feng, "Using semi-supervised clustering to improve regression test selection techniques," in Software Testing, Verification and Validation, 2011 IEEE Fourth International Conference on, 2011, pp. 1-10.

[24] M. Z. Z. Iqbal, Z. I. Malik, and M. Riebisch, "A model-based regression testing approach for evolving software systems with flexible tool support," in 2010 17th IEEE International Conference and Workshops on Engineering of Computer Based Systems, 2010, pp. 41-49.

[25] S. Mirarab, S. Akhlaghi, and L. Tahvildari, "Size-constrained regression test case selection using multicriteria optimization," IEEE Transactions on Software Engineering, vol. 38, pp. 936-956, 2012.

[26] Y. Pang, X. Xue, and A. S. Namin, "Identifying effective test cases through k-means clustering for enhancing regression testing," in Machine Learning and Applications (ICMLA), 2013 12th International Conference on, 2013, pp. 78-83.

[27] L. S. de Souza, R. B. Prudêncio, and F. d. A. Barros, "A Hybrid Binary Multi-objective Particle Swarm Optimization with Local Search for Test Case Selection," in Intelligent Systems (BRACIS), 2014 Brazilian Conference on, 2014, pp. 414-419.

[28] H. Hemmati and L. Briand, "An industrial investigation of similarity measures for model-based test case selection," in 2010 IEEE 21st International Symposium on Software Reliability Engineering, 2010, pp. 141-150.

[29] H. Cibulski and A. Yehudai, "Regression test selection techniques for test-driven development," in Software Testing, Verification and Validation Workshops (ICSTW), 2011 IEEE Fourth International Conference on, 2011, pp. 115-124.

[30] A. A. L. de Oliveira, C. G. Camilo-Junior, and A. M. Vincenzi, "A coevolutionary algorithm to automatic test case selection and mutant in mutation testing," in Evolutionary Computation (CEC), 2013 IEEE Congress on, 2013, pp. 829-836.

[31] L. S. de Souza, P. B. de Miranda, R. B. Prudencio, and F. d. A. Barros, "A multi-objective particle swarm optimization for test case selection based on functional requirements coverage and execution effort," in 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011, pp. 245-252.

[32] H. Hemmati, A. Arcuri, and L. Briand, "Empirical investigation of the effects of test suite properties on similarity-based test case selection," in 2011 Fourth IEEE International Conference on Software Testing, Verification and Validation, 2011, pp. 327-336.

[33] M. E. Delamaro and J. Offutt, "Assessing the influence of multiple test case selection on mutation experiments," in 2014 IEEE Seventh International Conference on Software Testing, Verification and Validation Workshops, 2014, pp. 171-175.

[34] J. Anderson, S. Salem, and H. Do, "Improving the effectiveness of test suite through mining historical data," in Proceedings of the 11th Working Conference on Mining Software Repositories, 2014, pp. 142-151.

[35] H. Hemmati, L. Briand, A. Arcuri, and S. Ali, "An enhanced test case selection approach for model-based testing: an industrial case study," Proceedings of the 8th ACM SIGSOFT international symposium on Foundations of software engineering, 2010, pp. 267-276.

[36] S. Huang, Y. Chen, J. Zhu, Z. J. Li, and H. F. Tan, "An optimized change-driven regression testing selection strategy for binary Java applications," in Proceedings of the 2009 ACM symposium on Applied Computing, 2009, pp. 558-565.

[37] H. Hemmati, A. Arcuri, and L. Briand, "Reducing the cost of model-based testing through test case diversity," in IFIP International Conference on Testing Software and Systems, 2010, pp. 63-78.

[38] S. Yoo and M. Harman, "Pareto efficient multi-objective test case selection," in Proceedings of the 2007 international symposium on Software testing and analysis, 2007, pp. 140-150.

[39] L. Yu, L. Xu, and W.-T. Tsai, "Time-constrained test selection for regression testing," in International Conference on Advanced Data Mining and Applications, 2010, pp. 221-232.

[40] A. Panichella, R. Oliveto, M. Di Penta, and A. De Lucia, "Improving multi-objective test case selection by injecting diversity in genetic algorithms," IEEE Transactions on Software Engineering, vol. 41, pp. 358-383, 2014.

[41] M. Kumar, A. Sharma, and R. Kumar, "Fuzzy entropy-based framework for multi-faceted test case classification and selection: an empirical study," IET software, vol. 8, pp. 103-112, 2013.

[42] Z. Xu, K. Gao, T. M. Khoshgoftaar, and N. Seliya, "System regression test planning with a fuzzy expert system," Information Sciences, vol. 259, pp. 532-543, 2014.

[43] E. Rogstad, L. Briand, and R. Torkar, "Test case selection for black-box regression testing of database applications," Information and Software Technology, vol. 55, pp. 1781-1795, 2013.

[44] L. S. De Souza, R. B. Prudêncio, F. d. A. Barros, and E. H. d. S. Aranha, "Search based constrained test case selection using execution effort," Expert systems with applications, vol. 40, pp. 4887-4896, 2013.

[45] B. Li, D. Qiu, H. Leung, and D. Wang, "Automatic test case selection for regression testing of composite service based on extensible BPEL flow graph," Journal of Systems and Software, vol. 85, pp. 1300-1324, 2012.

[46] Y.-D. Lin, C.-H. Chou, Y.-C. Lai, T.-Y. Huang, S. Chung, J.-T. Hung, et al., "Test coverage optimization for large code problems," Journal of Systems and Software, vol. 85, pp. 16-27, 2012.

[47] Z. Chen, Y. Duan, Z. Zhao, B. Xu, and J. Qian, "Using program slicing to improve the efficiency and effectiveness of cluster test selection," International Journal of Software Engineering and Knowledge Engineering, vol. 21, pp. 759-777, 2011.

[48] Y. Singh, A. Kaur, and B. Suri, "A hybrid approach for regression testing in interprocedural program," Journal of Information Processing Systems, vol. 6, pp. 21-32, 2010.

[49] N. Mansour, H. Takkoush, and A. Nehme, "UML-based regression testing for OO software," Journal of Software Maintenance and Evolution: Research and Practice, vol. 23, pp. 51-68, 2011.

[50] E. G. Cartaxo, P. D. Machado, and F. G. O. Neto, "On the use of a similarity function for test case selection in the context of model-based testing," Software Testing, Verification and Reliability, vol. 21, pp. 75-100, 2011.

[51] L. Zhang, "Hybrid regression test selection," in 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), 2018, pp. 199-209.B. Fu, S. Misailovic, and M. Gligoric, "Resurgence of Regression Test Selection for C++," in 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST), 2019, pp. 323-334.

[52] A. Labuschagne, L. Inozemtseva, and R. Holmes, "Measuring the cost of regression testing in practice: a study of Java projects using continuous integration," in Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, 2017, pp. 821-830.

[53] M. Vasic, Z. Parvez, A. Milicevic, and M. Gligoric, "File-level vs. module-level regression test selection for. net," in Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, 2017, pp. 848-853.

[54] A. Celik, M. Vasic, A. Milicevic, and M. Gligoric, "Regression test selection across JVM boundaries," in Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, 2017, pp. 809-820.

[55] B. Miranda and A. Bertolino, "Scope-aided test prioritization, selection and minimization for software reuse," Journal of Systems and Software, vol. 131, pp. 528-549, 2017.

[56] P. Kandil, S. Moussa, and N. Badr, "Cluster-based test cases prioritization and selection technique for agile regression testing," Journal of Software: Evolution and Process, vol. 29, p. e1794, 2017.

[57] B. Guo and M. Song, "Interactively decomposing composite changes to support code review and regression testing," in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 118-127.

[58] K. Wang, C. Zhu, A. Celik, J. Kim, D. Batory, and M. Gligoric, "Towards refactoring-aware regression test selection," in 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), 2018, pp. 233-244.

[59] Apache. (2017). JodaTime. Available: http://www.joda.org/joda-time/

[60] GitHub. Software code Hosting [Online]. Available: https://github.com/

[61] G. Rothermel and M. J. Harrold, "A framework for evaluating regression test selection techniques," in Software Engineering, 1994. Proceedings. ICSE-16., 16th International Conference on, 1994, pp. 201-210.

A Complexity Survey on Density based Spatial Clustering of Applications of Noise Clustering Algorithms

Boulchahoub Hassan[1], Rachiq Zineb[2], Labriji Amine[3], Labriji Elhoussine[4]
Laboratory of Systems Engineering (LaGeS), Hassania School of Public Works EHTP, Casablanca, Morocco[2]
Department of Mathematics and Computer Science, Faculty of Sciences Ben M'SIK, Casablanca, Morocco[1, 3, 4]

***Abstract*—Data Clustering is an interesting field of unsupervised learning that has been extensively used and discussed over several research papers and scientific studies. It handles several issues related to data analysis by grouping similar entities into the same set. Up to now, many algorithms were developed for clustering using several techniques including centroids, density and dendrograms approaches. We count nowadays more than 100 diverse algorithms and many enhancements for each algorithm. Therefore, data scientists still struggle to find the best clustering method to use among this diversity of techniques. In this paper we present a survey on DBSCAN algorithm and its enhancements with respect to time requirement. A significant comparison of DBSCAN versions is also illustrated in this paper to help data scientist make decisions about the best version of DBSCAN to use.**

Keywords*—Unsupervised learning; clustering; density clustering; DBSCAN

I. Introduction

The fast development of the internet and the availability of cheap mobiles, smart sensors and social networks applications allow users to generate a huge amount of data continuously. This rapid increase of data volume makes several domains difficult to be understood easily using only human capabilities. However many algorithms for clustering have been developed to guide data scientists to analyse and to understand data despite its volume. Nowadays, these algorithms play a crucial role in several sophisticated systems and applications including recommender systems, medical applications, face recognition, environmental assessment and anomalies detection [1][2][3][4][5]. To better understand any phenomena under investigation, clustering algorithms must extract correct and efficient statistics and trends, which is a very hard task, because results are often influenced by the nature of the real-world data which can be sparse, dense, spatial, high dimensional or even noisy. Therefore, algorithms must handle all complicated issues generated by data such as supporting volume increases, improving the scalability, processing high dimensional space, dealing with shaped structure and detecting outliers. The quality of clustering is also mainly influenced by the choice of the initial parameters such as number of clusters or the density radius. Thus, algorithms must vanish, optimize or even detect the parameters to use in order to detect meaningful clusters. To deal with all mentioned difficulties in real cases, many clustering approaches were raised including partitioning methods [6], hierarchical methods [7] and density based methods [8], etc.

In this paper, we are interested in density-based clustering, where clusters are defined by areas in which the density of the data points is high and clusters are separated from each other by areas of low density. We will focus especially on the DBSCAN algorithm [8] which can process spatial data efficiently and it can discard outliers properly. DBSCAN is a very simple and reliable technique, however it suffers from many limitations including its high complexity $O(n^2)$, its sensitivity to the local density variation, its dependence on initial parameters and its scalability failures. Therefore, it has undergone several improvements to make it efficient and to avoid its bastard chaos as effectively as possible. For instance, I-DBSCAN [9] and FDBSCAN [10] enhance the time requirement and minimize the deviation of results, MR-DBSCAN [11] improves scalability and deals with heavily skewed data and HDBSCAN [12] solves initial parameters issues, etc. We propose a comparison guide for all DBSCAN enhancements related to the complexity criterion and a repository for all DBSCAN versions related to time requirement is also presented in this work.

This paper is organized mainly in 4 sections: Sections 1 and 2 contains a brief refresh related to the clustering concept followed by an in-detail description about DBSCAN. Section 3 discusses the well-known DBSCAN improvements according to time complexity. Section 4 presents a comparison between DBSCAN versions based on time criteria.

II. Clustering Techniques

This section contains a brief description of partitional, hierarchical and density based clustering.

Clustering is the process of affecting each data object to a group based on the distance computation or on the similarity between each pair of observations. It is considered as the main process for many fields including image processing, pattern recognition, statistical data analysis and other business applications. Clustering methods can be broadly divided into several types including partitional, hierarchical, density based clustering, etc.

A. Partitional Clustering

Clustering has taken its roots from the partitioning method K-mean [6] which organizes all observations into an already

known number of groups (K). Each cluster is represented by its mean called centroid and objects are affected to the nearest cluster centroid. This method iterates many times over all observations to minimize the following objective function:

$$min \sum_{i=1}^{k} \sum_{x \in Si} ||x - \mu i||^2$$

Where Si is a Set of observations from a dataset, $Si \in \{S1, S2, \ldots, Sk\}$, k is the number of clusters and μi is the mean of points in Si.

K-mean is based on a very simple computation technique, however it is sensitive to outliers, data shapes and it assumes that clusters have roughly equal numbers of observations. In some cases, as mentioned in Fig. 1, it can lead to bad or even surprising results. Fig. 1(b), (f) and (d) are wrong clustering results.

The K-means method has relatively low time complexity and high computing efficiency, but it finds only compact and spherical shapes and it is still not suitable for non-convex data. Additionally, it needs prior knowledge about the number of clusters (K), it selects randomly the initial centroids. Thus, many improvements were done to overcome the limitations aforementioned such as the Partitioning Around Medoids (PAM) [13], Clustering Large Applications (CLARA) [14], and K-means for outlier detection [15]. Despite all efforts made, this type of clustering is not used when groups of data are expected to differ in size and shape, when the number of clusters is not known and when data contains noises. For instance, hierarchical and density clustering are explored to discover arbitrary shaped and meaningful clusters from large amounts of spatial data by preserving the spatial proximity of data objects.

B. Hierarchical Clustering

Hierarchical techniques seek to organize data objects into a tree structure representation called dendrogram. They are based on the computation of a symmetric distance matrix and they use some properly defined partitioning methods such as Ward's method [16], single or complete linkage. Several algorithms were invented under this type of clustering such as CURE [17] such as BIRCH [18].

As mentioned in Fig. 2, hierarchical techniques can be agglomerative or additive depending on where the algorithm starts processing the tree, from the top or from the bottom. Once the tree is built, hierarchical algorithms make splits in the additive processing or merge in agglomerative processing in order to find clusters. These cuts or merges decisions must be made properly thereby the quality of clusters will be better. For instance, as illustrated in Fig. 2, the level of cutting defines the number of clusters to detect. The first level gives rise to two clusters while the second one creates four clusters. Hierarchical algorithms are easy to understand and to implement. However, they rarely provide accurate results for mixed data types, they work poorly on very large data sets, they involve lots of arbitrary decisions and unfortunately, no adjustment can be performed once a merge or split decision has been executed. Many exceptions detected in partitioning and hierarchical clustering can be handled by using a density based clustering illustrated in the next section.

C. Density based Clustering

Density-Based Clustering refers to finding contiguous regions with high density among the dataset.

As mentioned in Fig. 3, these regions should be separated by low density regions called sparse regions. The idea behind such algorithms is that the clusters are represented by the detected dense regions and data objects in the sparse regions are typically considered noise/outliers.

In the next sections of this paper, we will focus on the most popular and the most cited density based algorithm (over 19430 times) called Density-Based Spatial Clustering of Applications with Noise [8].

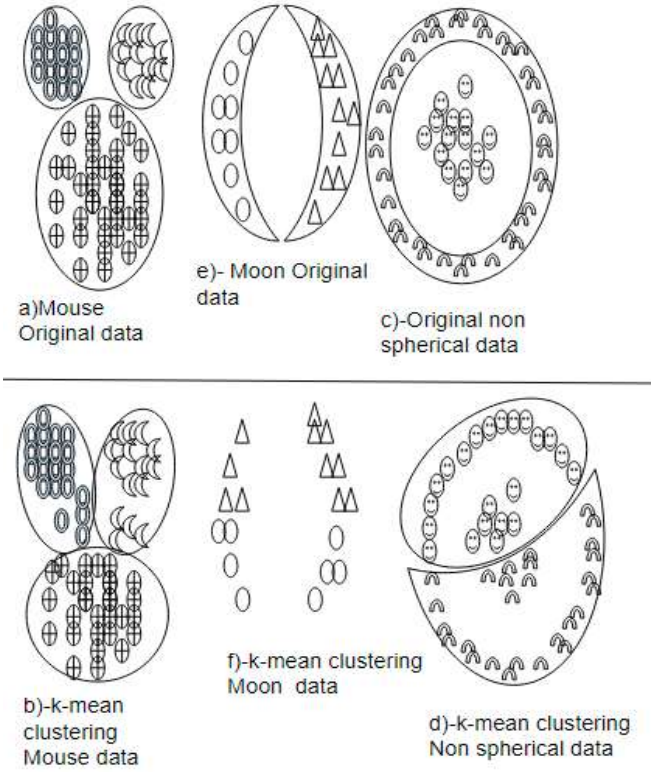


Fig. 1. K-Means Clustering Samples.

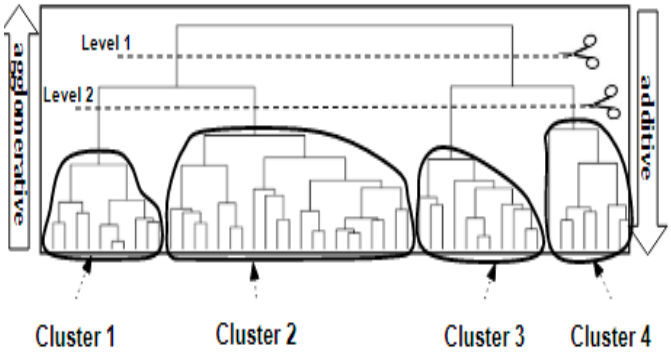


Fig. 2. Hierarchical Clustering Dendrogram.

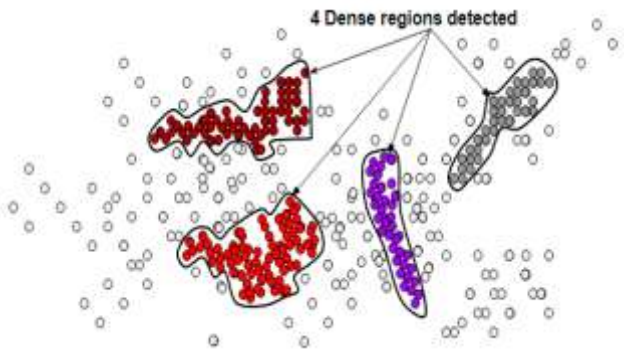


Fig. 3. Example for DBSCAN Clustering.

III. DBSCAN Basis

A. Definitions

DBSCAN received many scientific awards such as the test-of-time award from the leading data-mining conference KDD2014 for its good performance and its significant accuracy in clustering spatial data. The main purpose of DBSCAN is to detect arbitrary shaped clusters within a large data set and to effectively distinguish noises. It measure the density at any object O by counting the number of objects falling in a hyper sphere S(O,ε) where ε is a radius measured by an Euclidean distance. A region delimited by S(O,ε) is considered dense if the object O satisfies the following equation:

$$N\varepsilon\ (O) > MinPts\ where$$

$$N\varepsilon\ (O) = \{Q \in Dataset\ and\ distance\ (Q, O) < \varepsilon\}$$

$N\varepsilon\ (O)$ is the ε-neighbourhood of the object O and MinPts is the minimum number of points required to be present in the region to make hyper sphere S(O,ε) dense.

So, if objects share the same dense S(O,ε) then they belong to the same cluster. As mentioned before and to decide if a region is dense or sparse, this algorithm uses two parameters: an Euclidean distance threshold ε and a positive integer parameter MinPts.

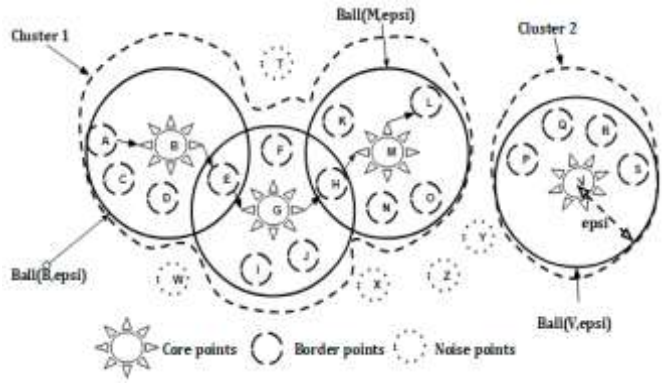


Fig. 4. .DBSCAN Core, Border and Noise Points.

As described below, DBSCAN introduces many definitions to categorize data objects into core, border or noisy objects [8].

DEFINITION 1: Core objects

An object O is considered core object if the number of objects inside the hyper sphere S(O,ε) is greater than MinPts parameter value. The points B, G, M are core objects in cluster 1 and V is a core object cluster 2.

DEFINITION 2: Border objects

An object is border if it belongs to some ε-neighbours of some core objects and the number of its own ε-neighbourhood is less than MinPts value. Thus, an object O is considered as a border object if it belongs to a cluster without being a core object. In Fig. 4, A,C,D,E,I,J,H,F,K,L,H,N and O are border objects for cluster 1 and P,Q,R and S are border objects for cluster 2.

DEFINITION 3: Noise objects

If an object O is not a core or border object then it is considered as a noise or outlier. In Fig. 4, the points T, W, X, Z, and Y are outliers.

DEFINITION 4: Directly Density reachability and Density reachability

If an object O is a core object, so all objects within the ε-neighbourhood of P are called directly density reachable objects from O. In Fig. 4, border objects A,C,D,E are directly reachable form the core point B and border objects P,Q,R,S are directly reachable from the core point V.

Two objects O1 and On are density reachable, if a chain of objects O1, O2, …, On is found within the dataset where Oi+1 is directly density-reachable from Oi with respect to the initial parameters ε, MinPts and $i\epsilon[1, n]$. For instance the chain A → B → E → G → H → M → L, in Fig. 4, makes a density link between the objects A and L. Thus A and L are density reachable.

DEFINITION 5: Maximality and Connectivity

Maximality: If a core object O belongs to a cluster, then all the objects density-reachable from O also belong to the same cluster.

Connectivity: If two objects O1, O2 belong to the same cluster so there is another object O in the same cluster such that both O1 and O2 are density-reachable from O. In Fig. 4, B and G are connected because they are density reachable through the chain B→E→G

B. DBSCAN Algorithm

Based on the previous definitions and the previously mentioned parameters ε and MinPts, we illustrate the DBSCAN algorithm in Table I.

TABLE I. DBSCAN ALGORITHM

DBSCAN ALGORITHM . DBSCAN(Data,ε,MinPts)

```
ClusterId = 0; // a cluster identifier
for each object O in Data do
   if O is not marked as "seen" then
      Mark O as "seen";
      Find Neighbors(ε,O,Data); // Neighbors of the object O using ε
param
      if  card(Neighbors(ε,O,Data))<MinPts then
          Mark O as "noise";
      else
          Mark O as "seen";
          ClusterId = ClusterId +1;
          Mark each object of  Neighbors(ε,O,Data) with cluster identifier
ClusterId;
          Add each object of Neighbors(ε,O,Data) which is not marked as
"seen" to the queue(ClusterId)
          while queue(ClusterId) is not empty do
               Take an object P from queue(ClusterId)  and mark it as
"seen"
               if card(Neighbors(ε,P,Data)>MinPts then
                   Mark each object of Neighbors(ε,P,Data) with cluster
identifier ClusterId;
                   if any object of Neighbors(ε,P,Data) is marked "noise"
then remove this mark.
                   Add each object of Neighbors(ε,P,Data) which is not
marked as "seen"  to
                   queue(ClusterId)
               end if
               Remove y from queue(ClusterId)
           end while
       end if
    end if
end for
Output all objects of Data along with their ClusterId or "noise" mark.
```

The previous algorithm describes DBSCAN where "Neighbors (ε, P, Data)" is the sub-set of objects in "Data" that are present in the hyper-sphere of radius at S(P, ε). "Card(Neighbors(ε,P,Data)" is the cardinality of the set "Neighbors(ε,P,Data)". Each object from "Data" is marked with a cluster identifier (ClusterId) which gives the cluster to which the object belongs or it is marked as "noise" indicating that the object is a noisy one. To distinguish between the objects which are processed from that which are not, the mark "seen" is used. Note that all objects of Neighbors(ε,P,Data) are initially marked as "noise", except the object P, then they can later become a border point of a cluster and hence the "noise" mark can be deleted.

According to the previous description, we can easily notice that DBSCAN does not require the pre-determination of the number of clusters and it requires only two parameters to determine when a region is considered to be dense or sparse, however it still suffers from several limitations including its high complexity which can reach $O(n\text{^}2)$, its failure with local density variation, its handicap related to data scalability and its huge memory consumption. Many works have been adopted to bring a significant optimization of the DBSCAN algorithm and to overcome its major drawbacks. For instance, E. Schubert et al. [19] discussed the relationship of the indexability of the dataset and the quality of clusters. They proposed some indicators of bad parameters to guide data scientists in choosing appropriate parameters ε and MinPts. For time reduction, B. Borah and D. K. Bhattacharyya used a sampling-based method [20] and J. Gan and Y. Tao used approximation techniques [21]. Derya Birant and Alp Kut tried to cover no spatial and spatial–temporal data by DBSCAN [22]. Moreover, other improvements were released to cluster in high dimensional space [23], to use parallel processing opportunities [24], [25] and to fix local density variation issues [26]. Knowing that complexity is a powerful criterion to decide about the efficiency of an algorithm, we propose a survey in the rest of this paper, a review of some well-cited DBSCAN extensions which significantly affect the time requirement.

IV. DBSCAN COMPLEXITY ENHANCEMENTS

Complexity criterion is among the ultimate indicators which qualify the efficiency of an algorithm. Thereby we decided to cover in this section some well-cited DBSCAN papers published between 2000 and 2019 and aiming to enhance the time requirement of the original algorithm. In the rest of this paper, "n" will represent the number of samples in the dataset and "d" will refer to the number of features studied. As mentioned in the first section, DBSCAN computes the empirical density for each dataset element and it measures mutual distances for the entire observations. Hence it requires a large volume of memory and a huge period of time to achieve large datasets clustering. Thus, it is qualified, by data scientists, as a very expensive algorithm and it is widely criticized due to its quadratic time requirement. Originally, Ester et al. [8] claimed that the DBSCAN will terminate in O (nlog(n)). However, the neighbourhood queries consume a big part of the running time. It requires $\sum_{p\in D}|N\varepsilon\,(p,D)| = O(n^2)$ to measure distances between all objects regardless of the initial parameters MinPts and ε. Fortunately, this time requirement can be reduced significantly to reach $O(n\,log_m n)$ [27] if some suitable indexing structure is used such as R*-tree [28] where m is the number of entries in a page of R*-tree. However, the use of R*-tree is suitable only when the dimensionality of the data is low. Thereby, researchers are still trying to run DBSCAN in some subquadratic time (i) by reducing the queries time and (ii) by minimizing the number of queries needed. As results of their efforts, many new methods appeared including hybrid methods [9] [29] which used only some accurate objects as prototypes rather than using all dataset objects. This approximation used by the hybrid methods can, in some cases, lead to clusters with bad quality. In the next paragraphs, we will weigh the pros and cons of some well cited DBSCAN time reducing methods.

B Borah et al. proposed IDBSCAN in 2004 to incorporate a sampling technique for searching the core object's neighbourhood. They used only outer objects as seeds and they ignored no representative objects. Therefore, they omit unnecessary queries by adding an extra function to the original algorithm based on Marked Boundary Objects (MBO) technique [20]. This function adds a complexity of O(sd) where s is the neighbourhood size. However, the overall complexity of IDBSCAN is $O(n\,log_m n)$ where m is the number of entries. Chen et al. [30] proposed an exact and approximate algorithm with $O(n^{(2-2/d+2)}\,polylog\,n)$ time for high dimensional space and $O(n^{1/2}\,polylog\,n)$ for two

dimensional spaces. P. Viswanath and Rajwala Pinkesh [9] proposed another fast hybrid density method called L-DBSCAN based on leaders clustering technique [31]. They derived some representative objects at the coarser level and others at the finer level of the clustering process. Authors used a first category of leaders to reduce the time requirement and second category to optimize the deviation of the results. This hybrid scheme uses only a set of pairs denoted by $\mathcal{L}^* = \{(\ell, \text{count}(\ell)) | \ell \in \mathcal{L}\}$ where ℓ is a leader and $\mathcal{L}$ a set of leaders. According to experiments showed by the authors, L-DBSCAN can run in $O((n+k)^2)$ where k represents the number of derived leaders which is much smaller than n. However, this technique can give raise to big margin error, when a leader is not originally dense but it is estimated dense according to $\mathcal{L}^*$. This method reduces the computation time, but it requires two additional thresholds: τc and τf.

P. Viswanath and V. Suresh Babu [29] enhanced the density approximation of leaders in their technique called rough-DBSCAN by combining the leaders clustering method [31] and the rough set approach [32]. They added a mapping between every leader and its belonging objects (followers). This mapping is represented by $\mathcal{L}^*$ = {(followers (l), count (l)} where l is a leader form $\mathcal{L}$. Then, they used a lower and upper approximation, as shown in equation 1, 2 and 3, to find the exact neighbours of a leaderN_ε (l, D).

$$N_{\varepsilon,\text{lower}-\tau}(l, D) \subseteq N_\varepsilon(l, D) \subseteq N_{\varepsilon,\text{upper}+\tau}(l, D) \quad (1)$$

$$N_{\varepsilon,\text{lower}-\tau}(l, D) = \cup_{l \in \mathcal{L}_1} \text{followers}(l) \quad (2)$$

$$N_{\varepsilon,\text{upper}+\tau}(l, D) = \cup_{l \in \underline{\mathcal{L}_1}} \text{followers}(l) \quad (3)$$

where $\mathcal{L}_1 = \{l_i \in \mathcal{L} \mid ||l_i - l|| < \varepsilon - \tau$

and $\underline{\mathcal{L}_1} = \{l_i \in \mathcal{L} \mid ||l_i - l|| \leq \varepsilon + \tau\}$

Rough-DBSCAN needs only $O(n+k^2)$ where k is the number of leaders, but it improves the clustering quality by minimizing the approximation error.

FDBSCAN [33] is another non-linear searching algorithm proposed by B. Liu, in 2007, to reduce redundant searching by using a fast merging algorithm. It sorts objects using dimensional coordinates and then selects only unlabelled objects outside a core object's neighbourhood in order to decrease region queries. Another interesting paper is proposed, in the same year, by Yi-Pu Wu et al. [34] to optimize the process of Nearest Neighbour Search by using Locality-Sensitive Hashing (LSH) technique. Authors used the hash collisions to detect and represent similarities between two objects A and B form a dataset D. On the other hand, to capture object similarity they compute the probability distributions over a set of hash functions $\mathcal{H}$ as $Pr_{h \in \mathcal{H}}[h(A) = h(B)] = S(A, B)$ where $h \in \mathfrak{H}$ and S is a similarity function defined as $S: DxD \rightarrow [0,1]$. This LSH algorithm makes a significant decrease in DBSCAN running time which becomes O (N) and maintains the quality of detected clusters [35].

Cheng-fa Tsa and Chien-Tsung Wu, inspired by the fast merging method of FDBSCAN [33], proposed the GF-DBSCAN algorithm to segment data into several grid-cells and to limit the neighbourhood searches only to the cell scope instead of exploring the entire grid. They merge clusters if they are intersected and the overlapping objects include some core object. By introducing this grid approach and this merging process, GF-DBSCAN minimizes significantly the number of searches and increases the clustering accuracy [36]. Gunawan [27] demonstrated that DBSCAN's performance can be improved to$O(n \log n)$ by applying the following process in order (i) partitioning data using a grid-cell (ii) determining all core points (iii) merging density-connected core points into clusters and finally (iiii) determining border points and noise. He used the hash table to discard cells without any point. However this faster algorithm is experimented only in two dimensional space. Therefore, J.Gan and Y.Tao extend Gunawan's thesis to $\mathbb{R}^d$and they get a running time of $\{O((nlogn)^{\frac{4}{3}}) if\ d = 3\}$ and $\{O(n^{\left(2-\frac{2}{\lceil \frac{d}{2} \rceil + 1} + \delta\right)}) if\ d >= 4\}$ where the parameter δ specifies the accuracy of the approximation. J.Gan and Y.Tao also proposed a new algorithm called ρ-approximate suitable for large datasets which can be computed in an expected time of $O\left(\frac{n}{\delta^{d-1}}\right)$. They are inspired by Chen et al.'s paper [30] which already discussed how to compute DBSCAN in $O(n \log n + n/\delta^{d-1})$.

GPUs opportunities and parallelization strategies are also used by some algorithms including G-DBSCAN algorithm [37] to speed up the original algorithm. G-DBSCAN constructs firstly a data graph G(O, E) where O are objects (nodes) connected by edges E if they are within a minimum proximity R (threshold parameter) of each other. Then it identifies clusters by using breadth-first search (BFS) technique [38]. Thereby, a complexity of O (n + ne) is added by the BFS search where 'ne' is the number of edges. G-DBSCAN uses graphics processing units (GPU's) capabilities to achieve acceleration greater than 100×, but unfortunately it doesn't reduce the original complexity.

The DBSCAN neighbour search operation can be optimized by using a graph-based index structure method, as demonstrated by K. Mahesh Kumar, and A. Rama Mohan Reddy [39]. Their idea is to prune out outliers objects early to vanish unnecessary distance computations which may be introduced by noises. RNN-DBSCAN [40] uses reverse nearest neighbour counts and k nearest neighbour graph traversals to estimate observation density. It reduces complexity of DBSCAN by using a single parameter (choice of k nearest neighbours) and also improves the ability of handling large variations in cluster density (heterogeneous density). Mark de Berg et al. [41] presented another O (n log n) approximate algorithm for DBSCAN in two dimensional space. They represented data objects using a smaller box graph $\mathcal{G}\_box$ where nodes are disjoint rectangular boxes with a diameter of at most ε and edges connect pairs of boxes within distance ε from each other. Then they detected another graph $\mathcal{G}\_core$ including only core points where the connected components of $\mathcal{G}\_core$ are considered as clusters. Mark de Berg et al. improved the quality of clusters by assigning borders to their nearest core point rather than the first cluster that finds them. Table II summarizes all aforementioned DBSCAN complexity enhancements.

TABLE II. TIME COMPLEXITIES OF THE WELL CITED DBSCAN VERSIONS

Year	Name	Time	Method used	Ref.
1996	DBSCAN	$O(n\hat{}2)$	Density based technique	[8]
		$O(n\log n)$	R*-tree	
2000	FDBSCAN: A fast DBSCAN algorithm	$O(n\log n)$	Representative points technique	[10]
2004	IDBSCAN	$O(n\ log_m n)$	Marked Boundary Objects (MBO)	[20]
2005	GEOMETRIC ALGORITHMS FOR DENSITY-BASED DATA CLUSTERING	$O\left(n^{2-\frac{2}{d+2}}\right)$ if d>=3 $O(n^{1.5}\ polylog\ n)$ for d=2	Computational geometry techniques. ε-fuzzy distance	[30]
2006	L-DBSCAN	$O((n+k)^2)$ K is the number of leaders	Leaders Clustering Method [31]	[9]
2006	FDBSCAN	$O(n\log n)$	Representative points technique kernel function	[33]
2007	A Linear DBSCAN Algorithm Based On LSH	$O(n)$	LSH : Locality sensitive hashing	[35]
2009	Rough DBSCAN	$O(n+k^2)$ K is the number of leaders	Set theory [32] leaders clustering method[31]	[29]
2009	GF-DBSCAN	1/100 of the time cost of FDBSCAN	Fast merging and Grid cells.	[36]
2010	TI-DBSCAN	Up to three orders of magnitude faster than DBSCAN.	Triangle Inequality property	[42]
2013	A faster algorithm for DBSCAN	$O(n\log n)$ for d=2	Grid partition Hash Table	[27]
2013	G-DBSCAN	Over than 100x faster than its sequential version using CPU	Data-graph (node and edges) breadth-first search BFS [38]	[37]
2015	ρ-approximate DBSCAN	$O((nlogn)^{\frac{4}{3}})$for d=3 $O(n^{2-\frac{2}{\lceil\frac{d}{2}\rceil+1}+\delta})$ for d>3 δ is a small positive constant	Approximation technique	[21]
2016	A fast DBSCAN clustering algorithm by accelerating neighbour searching using Groups method	improves the speed of DBSCAN by a factor of about 1.5–2.2	Graph-based method	[39]
2017	Faster DBSCAN and HDBCAN in Low-Dimensional Euclidean Spaces	$O(n\log n)$ $time\ for\ \mathbb{R}^2$	Box graph method	[41]

V. CONCLUSION

DBSCAN is a powerful technique for data clustering; however it still suffers from its huge time requirement which can reach $O(n^2)$ in the worst case. This paper weighs the pros and cons of the well-known and well-cited DBSCAN variations with respect to the time requirement. We present the current state of art related to DBSCAN complexity and also we mentioned some techniques used to enhance the original version of the algorithm. According to the papers studied, researchers can use the leaders clustering method, Graph-based method, breadth-first search BFS, Triangle Inequality Property or Locality sensitive hashing to bring new enhancements in this field. We noticed that DBSCAN complexity vary between $O(\mathrm{n})$ and $O(n^2)$. Another analysis of all these DBSCAN variations based on real data experiments will be presented in our further works.

ACKNOWLEDGMENT

Authors would like to express their special thanks of gratitude to Mr Labriji and Mr Rachik for their able guidance and support in completing this manuscript. We would also like to extend our gratitude to the anonymous reviewers whose thoughtful comments and suggestions will lead to improving this manuscript.

REFERENCES

[1] H. S. Emadi and S. M. Mazinani, "A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks," Wireless Personal Communications, vol. 98, no. 2. pp. 2025–2035, 2018, doi: 10.1007/s11277-017-4961-1.

[2] M. K. Najafabadi, M. N. Mahrin, S. Chuprat, and H. M. Sarkan, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," Computers in Human Behavior, vol. 67. pp. 113–128, 2017, doi: 10.1016/j.chb.2016.11.010.

[3] S. Khanmohammadi, N. Adibeig, and S. Shane Bandy, "An improved overlapping k-means clustering method for medical applications," Expert Systems with Applications, vol. 67. pp. 12–18, 2017, doi: 10.1016/j.eswa.2016.09.025.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015, doi: 10.1109/cvpr.2015.7298682.

[5] F. Yin and C.-L. Liu, "Handwritten Chinese text line segmentation by clustering with distance metric learning," Pattern Recognition, vol. 42, no. 12. pp. 3146–3157, 2009, doi: 10.1016/j.patcog.2008.12.013.

[6] J. B. MacQueen, "ON THE ASYMPTOTIC BEHAVIOR OF K-MEANS." 1965, doi: 10.21236/ad0629518.

[7] R. Sibson, "SLINK: An optimally efficient algorithm for the single-link cluster method," The Computer Journal, vol. 16, no. 1. pp. 30–34, 1973, doi: 10.1093/comjnl/16.1.30.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," 1996, Accessed: Dec. 08, 2020. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220.

[9] P. Viswanath and R. Pinkesh, "l-DBSCAN : A Fast Hybrid Density Based Clustering Method," 18th International Conference on Pattern Recognition (ICPR 06). 2006, doi: 10.1109/icpr.2006.741.

[10] Z. Shui, "FDBSCAN: A Fast DBSCAN Algorithm," 2000, Accessed: Dec. 15, 2020. [Online]. Available: https://www.semanticscholar.org/paper/FDBSCAN%3A-A-Fast-DBSCAN-Algorithm-Shui/d6d1e7e468035b63138d6a4de4ca5685fb700808.

[11] Y. He, H. Tan, W. Luo, S. Feng, and J. Fan, "MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data,"

Frontiers of Computer Science, vol. 8, no. 1. pp. 83–99, 2014, doi: 10.1007/s11704-013-3158-3.

[12] R. J. G. B. Campello, Ricardo J G, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," Advances in Knowledge Discovery and Data Mining. pp. 160–172, 2013, doi: 10.1007/978-3-642-37456-2_14.

[13] "Kaufman, L. and Rousseeuw, P.J. (1990) Partitioning around Medoids (Program PAM). In Kaufman, L. and Rousseeuw, P.J., Eds., Finding Groups in Data An Introduction to Cluster Analysis, John Wiley & Sons, Inc., Hoboken, 68-125. - References - Scientific Research Publishing." https://www.scirp.org/(S(czeh2tfqyw2orz553k1w0r45))/reference/ReferencesPapers.aspx?ReferenceID=1771062 (accessed Dec. 04, 2020).

[14] L. Kaufman and P. Rousseeuw, "Clustering Large Applications (Program CLARA)," 2008, doi: 10.1002/9780470316801.CH3.

[15] S. Chawla and A. Gionis, "k-means-: A Unified Approach to Clustering and Outlier Detection," 2013, Accessed: Jan. 06, 2021. [Online]. Available: https://pdfs.semanticscholar.org/70f4/5be50599f12a1b682a192c3c48ebda0bb1c4.pdf.

[16] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, vol. 58, no. 301. pp. 236–244, 1963, doi: 10.1080/01621459.1963.10500845.

[17] S. Guha, R. Rastogi, and K. Shim, "Cure: an efficient clustering algorithm for large databases," Information Systems, vol. 26, no. 1. pp. 35–58, 2001, doi: 10.1016/s0306-4379(01)00008-4.

[18] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," 1996, Accessed: Jan. 03, 2021. [Online]. Available: http://dl.acm.org/citation.cfm?id=233324.

[19] E. Schubert, J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," ACM transactions on database systems, vol. 42, no. 3, p. 6, 2017, Accessed: Dec. 15, 2020. [Online].

[20] B. Borah and D. K. Bhattacharyya, "An improved sampling-based DBSCAN for large spatial databases," International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of. doi: 10.1109/icisip.2004.1287631.

[21] J. Gan and Y. Tao, "DBSCAN Revisited: Mis-Claim, Un-Fixability, and Approximation," 2015, Accessed: Dec. 17, 2020. [Online]. Available: http://dl.acm.org/citation.cfm?id=2737792.

[22] "ST-DBSCAN: An algorithm for clustering spatial–temporal data," Data Knowl. Eng., vol. 60, no. 1, pp. 208–221, Jan. 2007, Accessed: Dec. 08, 2020. [Online].

[23] L. Ertöz, M. Steinbach, and V. Kumar, "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data," Proceedings of the 2003 SIAM International Conference on Data Mining. 2003, doi: 10.1137/1.9781611972733.5.

[24] G. Liu, B. Qiu, and L. Wenyin, "Automatic Detection of Phishing Target from Phishing Webpage," 2010 20th International Conference on Pattern Recognition. 2010, doi: 10.1109/icpr.2010.1010.

[25] Y. He et al., "MR-DBSCAN: An Efficient Parallel Density-Based Clustering Algorithm Using MapReduce," 2011 IEEE 17th International Conference on Parallel and Distributed Systems. 2011, doi: 10.1109/icpads.2011.83.

[26] P. Liu, D. Zhou, and N. Wu, "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise," 2007 International Conference on Service Systems and Service Management. 2007, doi: 10.1109/icsssm.2007.4280175.

[27] A. Gunawan, "A faster algorithm for DBSCAN," 2013, Accessed: Dec. 28, 2020. [Online]. Available: https://www.semanticscholar.org/paper/A-faster-algorithm-for-DBSCAN-Gunawan/138f3e2aac21ca81fb7bf093ebae07859111e6dd.

[28] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: an efficient and robust access method for points and rectangles," Proceedings of the 1990 ACM SIGMOD international conference on Management of data - SIGMOD '90. 1990, doi: 10.1145/93597.98741.

[29] P. Viswanath and V. Suresh Babu, "Rough-DBSCAN: A fast hybrid density based clustering method for large data sets," Pattern Recognition Letters, vol. 30, no. 16. pp. 1477–1488, 2009, doi: 10.1016/j.patrec.2009.08.008.

[30] D. Z. Chen, M. Smid, and B. Xu, "GEOMETRIC ALGORITHMS FOR DENSITY-BASED DATA CLUSTERING," International Journal of Computational Geometry & Applications, vol. 15, no. 03. pp. 239–260, 2005, doi: 10.1142/s0218195905001683.

[31] H. Spaeth, "Cluster analysis algorithms for data reduction and classification of objects," 1980, Accessed: Jan. 08, 2021. [Online]. Available: https://cds.cern.ch/record/102044.

[32] Z. Pawlak, "Rough sets," International Journal of Computer & Information Sciences, vol. 11, no. 5, pp. 341–356, Oct. 1982, Accessed: Dec. 29, 2020. [Online].

[33] B. Liu, "A Fast Density-Based Clustering Algorithm for Large Databases," 2006 International Conference on Machine Learning and Cybernetics. 2006, doi: 10.1109/icmlc.2006.258531.

[34] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," 2004, Accessed: Jan. 12, 2021. [Online]. Available: http://dl.acm.org/citation.cfm?id=997857.

[35] Y.-P. Wu, J.-J. Guo, and X.-J. Zhang, "A Linear DBSCAN Algorithm Based on LSH," 2007 International Conference on Machine Learning and Cybernetics. 2007, doi: 10.1109/icmlc.2007.4370588.

[36] C.-F. Tsai and C.-T. Wu, "GF-DBSCAN: a new efficient and effective data clustering technique for large databases," 2009, Accessed: Dec. 15, 2020. [Online]. Available: https://www.semanticscholar.org/paper/GF-DBSCAN%3A-a-new-efficient-and-effective-data-for-Tsai-Wu/909e93cbf1867e2b5af089810bdbb8352e75ff53.

[37] G. Andrade, G. Ramos, D. Madeira, R. Sachetto, R. Ferreira, and L. Rocha, "G-DBSCAN: A GPU Accelerated Algorithm for Density-based Clustering," Procedia Computer Science, vol. 18. pp. 369–378, 2013, doi: 10.1016/j.procs.2013.05.200.

[38] E. F. Moore, The Shortest Path Through a Maze. 1959.

[39] K. M. Kumar, K. Mahesh Kumar, and A. Rama Mohan Reddy, "A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method," Pattern Recognition, vol. 58. pp. 39–48, 2016, doi: 10.1016/j.patcog.2016.03.008.

[40] A. Bryant and K. Cios, "RNN-DBSCAN: A Density-Based Clustering Algorithm Using Reverse Nearest Neighbor Density Estimates," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 6. pp. 1109–1121, 2018, doi: 10.1109/tkde.2017.2787640.

[41] M. de Berg, A. Gunawan, and M. Roeloffzen, "Faster DB-scan and HDB-scan in Low-Dimensional Euclidean Spaces," Feb. 28, 2017.

[42] M. Kryszkiewicz and P. Lasek, "TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality," in Rough Sets and Current Trends in Computing, Jun. 2010, pp. 60–69, Accessed: Jan. 01, 2021. [Online].

Particle Physics Simulator for Scientific Education using Augmented Reality

Hasnain Hyder[1], Gulsher Baloch[2], Khawaja Saad[3], Nehal Shaikh[4], Abdul Baseer Buriro[5], Junaid Bhatti[6]
Department of Electrical Engineering
Sukkur IBA university

***Abstract*—In this era of fourth industrial revolution, young learners need to be equipped with 21st century skills, such as critical thinking, creativity, communication, collaboration, innovation and problem solving. Augmented Reality (AR) based learning systems are an effective tool to embed these skills. This paper presents a detailed review of latest research on an AR-based learning systems. Furthermore, an AR-based learning system is proposed to demonstrate the particle physics experiments i.e. proton-proton collision and Higgs field. The proposed learning system algorithms are developed using particle system of unity 3D software. Then, Microsoft Kinect sensor is interfaced with unity 3D to create an immersive experience. Then, the qualitative analysis of the proposed system and latest AR-based learning systems is presented. Finally, the quantitative analysis of the proposed system is conducted. Overall, the results suggest that 85% of the participants recommended the proposed learning system.**

Keywords*—Particle physics; augmented reality; proton-proton collision; Higgs field; interactive classroom; AR in education; AR based lab experiments

I. Introduction

According to European Union, AR will be one of the emerging technologies to pave the way for the development of smart industry in near future [1], [2]. The disruptive technologies are extensively utilized in many applications to enhance their performance. However, many sectors are still lacking in adapting the latest technology. Education sector is considered as one of them [3]. In the traditional form of learning, the teacher delivers knowledge while students act as recipients only. However, students find interactive way of learning to be more exciting and effective. AR have a key role in developing learning systems to make the learning process more effective and less tedious [4]. It enables human-machine interaction while overlaying virtual components on real world environment. It has potential applications in multiple fields such as education, health-care, rehabilitation, etc. [5]–[7]. It is combined with holographic technology to create applications for museums and other visiting places to display art and culture [8], [9]. It helps to create new and more effective learning systems to develop critical thinking, creativity, communication, collaboration, innovation and problem solving.

The concept of AR was introduced in the early 90s. Since then, significant advancements have been made in this field [10]. These advancements have created multiple opportunities to develop systems and products that provide immersive experience to the users'. This technology provides novelty by combining real and virtual world and registering it to 3D reality [11]. The advancement in AR has introduced new teaching methods which enhances the interest and class participation of students [12]–[17]. Recent studies show that students engaging in such interactive ways of learning have great positive impact on their education [14].

Recent studies proposed by Latvian teachers tells us that old methods of teachings have reduced the interest and concentration level of the students [15]. The overall results in Fig. 1 shows that the majority of students have decreased interest in the current methods of teachings [15].

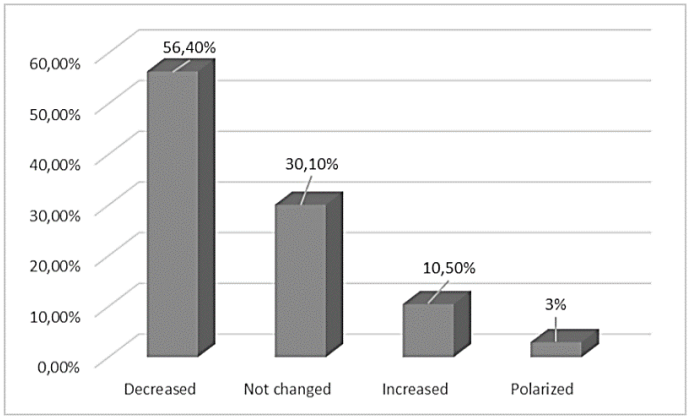


Fig. 1. Opinion of Teachers about Changes of Students' Learning Motivation [15].

In addition, some studies have been conducted to assess the usability of augmented reality in educational field. In one of the study, D.L Hakim et al. [10] concluded that AR has a significant impact as a learning media and has greatly affected students' motivation and learning outcomes. Annafi A et al. in [10] reviewed and collected data from 30 articles published in the past 10 years. The articles reviewed the cognitive, affective, and psychomotor aspects that include students' cognitive skills, learning abilities, understandings, motivation, responses, attentiveness, involvement, and outcomes towards any study material [10].

In another review, Nicholas Pellas et al. [16] analyzed the literature on AR with Game-based Leaning (ARGBL) approach. The ARGBL system was developed while keeping in view the advantages, disadvantages, instructional affordance and effectiveness across various primary and secondary education [16]. A detailed methodology proposed by Kitchenham's paper published in 2007 was adapted for the purpose of systematic review [16]. The journals were selected through inclusive and exclusive criteria and then the data was catego-

rized and analyzed. It was concluded that the ARGBL usage has significantly increased for teaching Science, Technology, Engineering and Mathematics (STEM) in the past few years.

In another review, Marina Ismail et al. [18] provided a comprehensive review of existing studies on the use of Kinect device in education and rehabilitation. A total of 16 studies were collected, analyzed and organized in a detailed order [18]. It concluded that Kinect-based systems are beneficial in providing e-learning environment and interactive experiments [18]. The outcome of these reviews demonstrates an overall positive impression and serves as a motivational factor for further analysis on AR as a learning tool. Similarly, in this paper a systematic review of 10 research papers is carried out focusing on AR-based educational systems.

Our Contributions are as follows: *i*) A detailed review of the latest AR-based learning systems. *ii*) Developed an algorithms to simulate the particle physics experiments i.e. proton-proton collision and Higgs field. *iii*) Designed an AR-based learning system to demonstrate the particle physics experiments with an immersive experience.

II. Literature Review

In this section, a detailed review of recent research papers is conducted to show the impact of AR on modernizing the classroom learning. The goal of this review is to lay out a comprehensive analysis about the impact and findings of recently proposed systems in the field of education. The steps taken to select the qualitative research papers are as follows:

1) The data is collected from academic journals of MDPI, IEEE ACCESS, Springer, International journal of Engineering and Technology, Elsevier, Canadian center of Science and Education, International journal of Geographical Information Science, Journal of Physics, and Hindawi which are published in the years from 2015 through 2020.

2) Numerous keywords such as "Kinect sensor and AR", "AR in education", "Impact of AR in educational field", "Lab based experiments on Kinect sensor", "Interactive wall and floor in classrooms", "Advantages of using AR in classrooms", "AR effects on children learning" were used to find the most relevant articles.

3) Approximately 40 articles were selected and analyzed keeping in view the title, abstract and keywords of the paper. Total of 11 studies were identified as the most relevant to the topic and hence were extensively reviewed.

The summary-based analysis of each study is given below:

- Franca Gorzotto et al. [19] created a Magic Room using projector, Kinect sensor, several smart objects and Unity 3D software for children with Neural Developmental Disorder (NDD). The games developed, in this paper, detect the children's behavior as they interact with the multimedia content in the Magic Room [19]. The experiments proved to elicit functional performances, social behaviors, and emotional responses. The authors concluded that further empirical research is needed in this area as the experiment was limited to health care [19], [20].
- Plamen D. Petrov et al. [12] analyzed the effect of AR on students' learning performance in STEM education. This experiment uses ZSpace which is an all-in-one AR system comprising of virtual reality monitor and a computer. It combines AR and VR to create an immersive and interactive experience as shown in Fig. 2 [12]. The experiment was carried out on 80 participants and a significant difference was observed in students' understanding as compared to the traditional system [12]. In conclusion ZSpace introduces high level of personalization and helps improve the understanding of students. It allowed students to explore and practice without worrying about financial (supporting lab equipment) or ethical (animal injury in biology lab) issues [12].

Fig. 2. ZSpace as an AR Tool for STEM Education [12].

- Lidice Haz et al. [3] implemented a Kinect-based multimedia system for children of primary schools for increasing classroom participation. This system was developed in a cascaded model consisting of level-based design with incrementing difficulty. The models enabled students to add, subtract, complete words and complete sentences. A survey indicated that 87% of students preferred this method of learning [3].
- Nak-Jun Sung et al. [21] investigated the applications of the Physics using AR. A video see-through method is used to construct an AR environment by using Kinect V2 sensor. The experiment uses soft body simulator version MSS (Mass Spring System) because of its high simulation accuracy and speed. The AR experiment first combines a real time video stream with a soft body simulation as shown in Fig. 3. Then, it creates several objects with various material properties by changing the object motion through simulation as shown in Fig. 4. In conclusion, a survey showed that 93% of responders were in favor of teaching the Physics using this simulator [21].
- Tamas Matuszka et al. [22] developed a gesture-controlled educational gaming system. Authors reduced the cost of the system by using deep learning method. The algorithm first detects the object and uses "sliding window method" gesture collection. Keras and Tensorflow were used as deep learning backend. In conclusion, ordinary camera with the proposed algorithm provided similar results as depth-camera based system.

Fig. 3. Combined System of Soft Body Simulator and Video Stream obtained from Kinect Device [21].

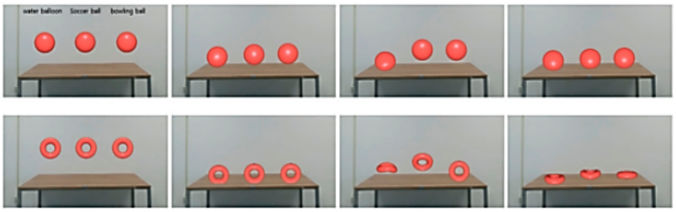


Fig. 4. Result of Simulation of 1st Scene Comprising of Sphere Model [21].

- Raul Lozada-Yanez et al. [23] designed a Kinect-based AR Math Learning System (KARMLS) for increasing student performance in mathematics. It involved 29 third-grade children from Riombamba city, Ecuador. The system comprises of sumar (addition), ordeanr (arranging shapes in order) and parear (making pairs). The Fig. 5 shows student interaction with KARMLS using wave gesture. It was concluded that the system had a positive impact on students and was more effective on low grades securing students [23].

Fig. 5. Natural Interaction between End User and Kinect based Prototype [23].

- Corey Pittman et al. [24] explored the utility of AR for knowledge of the Physics in the classroom. A PhyAR prototype was developed using Unity3D and mixed reality toolkits with Microsoft HoloLens [24]. Coulomb's Law, elastic collision, parallel circuits, volume etc. were demonstrated using virtual objects in the physical space as shown in Fig. 6. Fifteen participants were gathered to explore each concept using prototype and then fill a questionnaire for feedback [24]. The students' response was positive but HoloLens being a head worn device restricted the physical world interaction. Feedback from the students emphasized on adding interaction with real world objects [24].

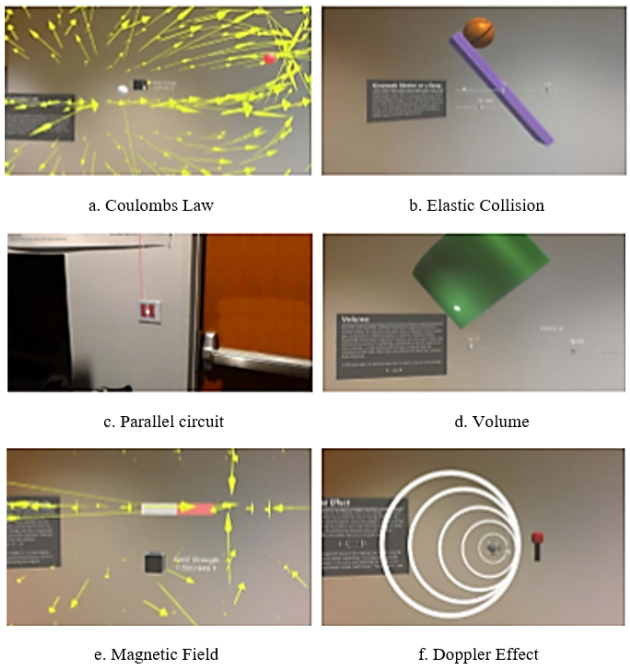


Fig. 6. Presentation of Some Physics Topics shown on PhyAR Application on Microsoft Hololens [24].

- Mingshao Zhan et al. [25] studied the recent developments in game based virtual educational laboratories using Kinect sensor. Kinect device is used to scan the real-world data and a map is created to implement a virtual laboratory. The RANdom SAmple Consensus (RANSAC) algorithm is used for shape detection as shown in Fig. 7 [25]. In this way, integration of real materials with natural materials have been observed and presented [25], [26]. Microsoft Kinect sensor proved to be realistic and affordable for virtual environment modeling, human-computer interface, and hardware interface implementation.

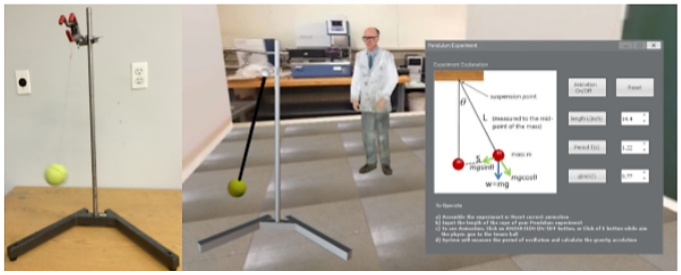


Fig. 7. Experimental Setup of Foucault Pendulum (Left); Foucault Pendulum Experiment Implementation in Game-based VE (Right) [25].

- Mingliang Xu et al. [27] provided a Kinect-based system for physical education of trainees (children) without trainers. For training this system, Hierarchical

Hidden Markov (HMM) based algorithm was used. It allowed trainers to develop customized training paths for each individual trainee. This method significantly enhanced the effects of physical training in the absence of trainers [27].

- Yi-Hsing Chang et al. developed a Kinect-based English learning system [28]. They integrated Kinect as the interaction technique with theories of situated learning and attention, relevance, confidence, and satisfaction (ARCS) model [28]. This system enables to plan and design the learning activities as per situated learning [28]. The system provides virtual environment which helps to achieve spatial and physical experience, assisting learner's engagement, and enhancing learning motivation as shown in Fig. 8. The authors concluded proposed system improved students learning motivation [28].

Fig. 8. English Learning system using ARCS Model) [28].

- Maria Cristina Costa et al. presented a mobile augmented reality based application called PlanetarySystemGo which is a location-based game to promote learning about the universe [29]. The architecture of the system is divided into three components: platform server (manages all data in the system), Web application (the assessment of learning outcomes with a back office) and mobile app (created using unity 3D to provide dynamic environment to create AR content) as shown in Fig. 9. Several surveys including questionnaires were conducted with primary school students and teachers. According to the results, the application enhances the students' interest to learn about solar system and keep them engaged [29].

The above studies are further analyzed and categorized into major contribution and limitation provided by authors. It provides all-inclusive information about how AR with the help of Kinect sensor has changed the way of learning. Data is taken exclusively from the said articles and has been compared in Table I.

III. Proposed System Overview

This section presents the development of proposed AR-based learning system. The system is developed to demonstrate

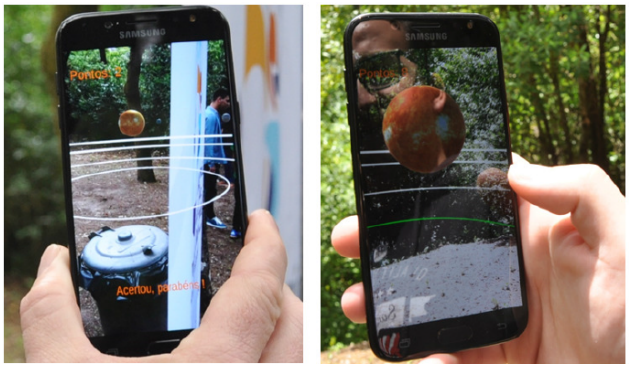


Fig. 9. Mobile App Showing Hunting Orbits and Planets.) [29].

the experiments of particle physics. The proposed system is developed to simulate the proton-proton collision and Higgs field. The algorithms are developed using development software Unity 3D, Microsoft Kinect V2 , a projector, C# (C-sharp) language and Visual Studio IDE. Kinect sensor was interfaced with Unity 3D to enable interaction of objects with the environment. The Microsoft SDK for Kinect sensor is used to track the body with the help of its infrared-based depth camera. It accurately captures the real time 3D scene and generates the built-in skeletal using customized software [30], [31].

In the proposed system the actions performed by human body are acquired and processed by the Kinect Sensor. Then these actions are evaluated on basis of the developed algorithms and resultant output is delivered via Projector. The proposed system is shown in Fig. 10.

The developed environment is projected on the floor providing the real-time interactive learning experience. The algorithm developed for Higgs field allows the user to visualize their reflection inside an particles filled environment.

Fig. 10. The Proposed System.

Algorithms are developed to demonstrate the particle behavior such as proton-proton collision effect and Higgs field.

Proton collision algorithm shows the generation of new particles as a result of proton-proton collision. In this algorithm, the environment is designed where students can play a football game in which the footballs are supposed as protons.

IV. Proposed System Implementation

A. Proposed Algorithm for Proton-Proton collision

The collision of protons is possible through LHC (Large Hadron Collider) tunnel, which is particularly a particle accelerator owned by CERN laboratories [32]. This tunnel allows

TABLE I. DETAILS ABOUT STUDIES RELATED TO AR TECHNOLOGY IN EDUCATIONAL FIELD.

Source	Major Contribution	Limitations provided by authors
[19]	Magic room developed in this paper has a strong potential as a learning environment for children with Neurodevelopmental Disorder (NDD)	Regular update of the content in areas like communication, Psychomotor, emotion, and cognition is required
[12]	The data obtained from performing experiments revealed that the integration of AR allows students to explore, practice and interacts with Science, Technology Engineering and Mathematics (STEM) content with an effective way.	Price of the overall system is 7000 dollar which makes it costly
[3]	It concludes that Kinect sensor with multimedia technology facilitates the teaching and learning process through an attractive and motivating environment	Physical interaction should be enabled for more enhanced and learning experience.
[21]	According to the findings, the proposed simulator helps to teach the Physics effectively due to more realistic representation of complex processes	Requires more realistic representation of certain content and high specification. Additional equipment is required to expand the project.
[22]	It has been found that a cost-effective monocular camera-based gesture recognition method can ensure similar level of recognition accuracy as depth-camera based solutions	Limited gesture recognition as compared to Microsoft Kinect.
[23]	The outcome of experiments suggested that there is a significant improvement in grades of students using Kinect Based AR Math Learning System (KARLMS)	Better visual graphics and interactivity should be enabled for higher accuracy.
[24]	The participants' evaluation revealed that the users desired to see such 3D AR content in the Physics.	Limited field of view, integration of physical objects, and support of environment around the user are required.
[25]	According to findings of the research, Microsoft Kinect based educational VR laboratory proved to be efficient, realistic, and affordable as compared to traditional approaches	Integration of more tracking algorithms into DAQ software package is required.
[27]	User study evaluation contributes to the research that the effect of Kinect-based training method is much better than the traditional video-based method.	Better visual graphics and interactivity should be enabled for higher accuracy.
[28]	This system helped students to learn English effectively through integrating Kinect with situated learning and ARCS model.	The system seizes to operate during experiment. Furthermore, instructions were too descriptive having small font size, making them unreadable.
[29]	it has been noted that the PlanetarySystemGo platform has great potential in serving as an informal and formal learning environment about the solar system for all students.	Some technological hindrances were observed such as instability due to GPS coordinates and inaccurate gyroscope reading. Also, the information content should be more enhanced and upgraded such as more planetary system need to be introduced.

the particles to travel approximately at the speed of light (3 meters per second less than the speed of light), and gain higher energies. The number of particles dispersed after the proton-proton collision is directly dependent upon the speed and energy of the proton. The scientists observe the new particles by colliding protons at a higher velocity. The new particles are smaller as compared to protons and have different properties altogether [32].

The proposed system simulates the LHC tunnel to visualize the generation of new particles after collision of protons. This system is developed in the form of a game to enable the users' interaction in the particle acceleration, collision and generation in a fun and interesting way. Using this system students can observe that the amount of new particles generated is directly proportional to the velocity of protons.

The proposed particle behavior system follows the String Theory / String Model. It states that, upon collision, two protons divide into particles resembling a string [33]. The distribution of the particles according to String model follows the Equation (1):

$$\mathbf{x}_1 = (1 - y)\mathbf{x}_o \tag{1}$$

where, $\mathbf{x}_1$ represents the number of particles created after collision, $\mathbf{x}_o$ is the energy applied by the kick, whereas y is the fraction of energy carried by string (distributed particles). The value of y exists between 0 and 1 depending upon the distribution function. Equation (1) can be further extended for the number of new particles.

$$\mathbf{x}_n = (1 - y_{av})^i \mathbf{x}_o \tag{2}$$

Where $\mathbf{x}_n$ is the residual energy of the string after i^{th} rank of distribution, y_{av} is the average value of y, and i is the average multiplicity of the produced particles which can be calculated using Equation (3) and Equation (4).

$$i = \ln \frac{\mathbf{x}_o}{\mathbf{x}_n} \beta \tag{3}$$

where,

$$\beta = \frac{-1}{\ln(1 - y_{av})} \tag{4}$$

The proposed simulated environment allows two users to kick the protons (simulated as footballs) in a real-time. The momentum of the kick determines the speed of protons. Users can visualize the generation of new particles when protons collide.

Taking our proposed system into consideration, the external force applied (kick) is responsible for increasing the length of the string/number of distributed particles. The higher the force, the more the production of particles.

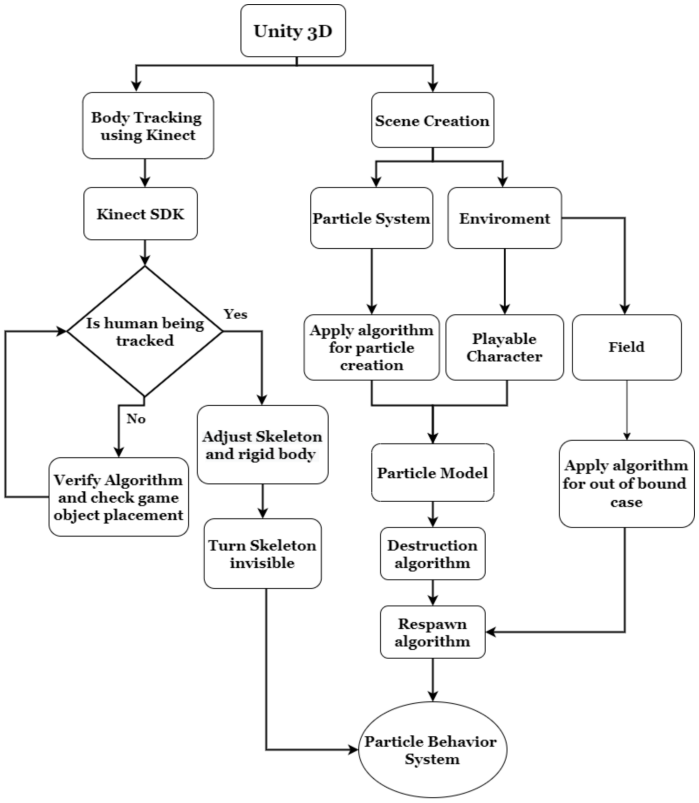


Fig. 11. Particle Behavior System Flow chart.

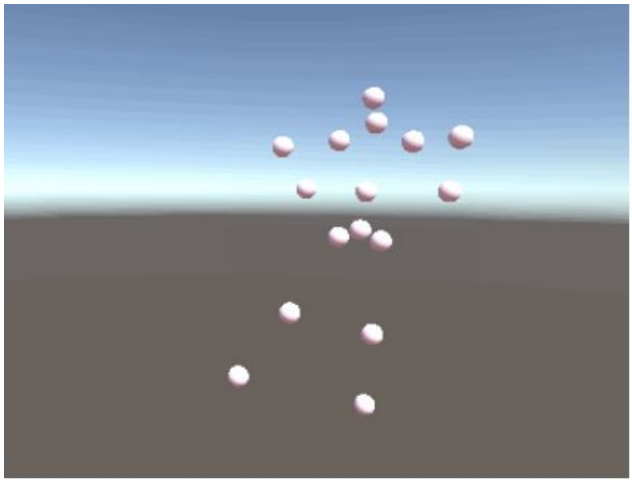


Fig. 12. Skeleton Tracking using Kinect.

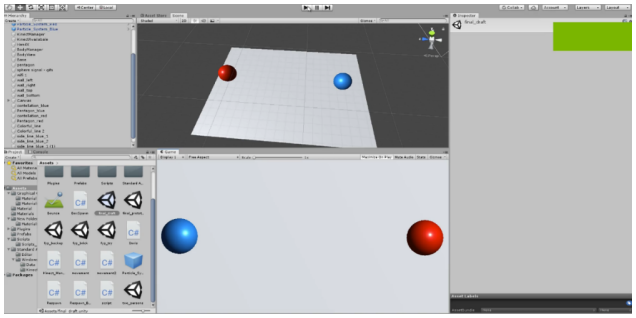


Fig. 13. Two Spheres Representing Two Protons.

1) Implementation of Proton-Proton Collision: Firstly, we created an empty 3D project inside Unity 3D, allowing us to develop a system in 3D environment. After that Microsoft Kinect V2 was connected to PC via adapter and its SDK was imported inside Unity. Methods available in SDK were used to create scripts that allowed human joints tracking. The joints of tracked bodies were created using 3D objects like spheres, as displayed in Fig. 12. These objects were also provided with rigid body and collider modules through scripting. Rigid body module allows the object to have properties like gravity, force, acceleration etc. Whereas collider module allows the object to collide with any other surface. If the collider module is not enabled, the objects cannot collide and pass through one another.

A scene was created in which the system was to be developed. Firstly, the environment was developed using 3D objects. The playable characters (Protons) were created using two Spheres and the Field (platform) was designed using multiple cubes arranged accordingly as displayed in Fig. 13. Playable characters were provided with rigid body and collider modules, whereas only collider module was provided to the field. A script was attached to the field objects so that the spheres return to their original position whenever they were out of bounds.

The built-in particle generation system of Unity 3D was used for particle generation as shown in Fig. 14. This system helps generate particles of various sizes and quantities. It also helps to create trails behind each particle. A C# script is used to simulate the number of particles generated using the speed of spheres, using the Equation (2) of String Model.

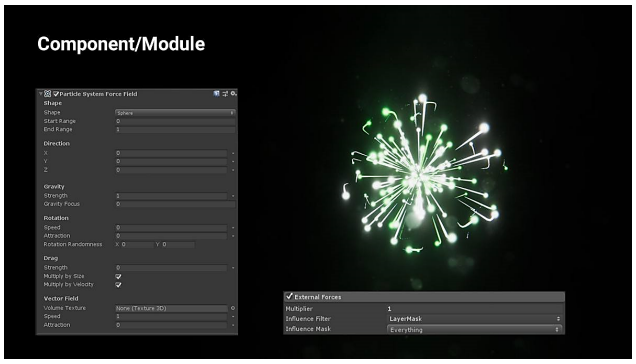


Fig. 14. Unity's Particle Generation System.

Graphics were embedded on each of the game object and were created using Adobe Photoshop and Adobe Premiere Pro as displayed in Fig. 15. The system was named "Particle Model" and was provided with scripts that allowed the destruction of spheres upon collision. This script allowed the spheres to be destroyed upon collision and generates particles as shown in Fig. 16. Another script was attached to the model that allowed the particles to respawn at original positions. Speed bar is included in the environment to observe variation in speed of the spheres. Particle model is finalized by overlapping tracked bodies over the particle model while converting the skeleton

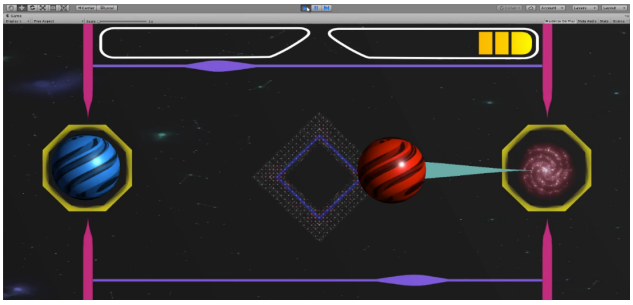


Fig. 15. Graphical Interface of Proton-Proton Collision System on Unity 3D.

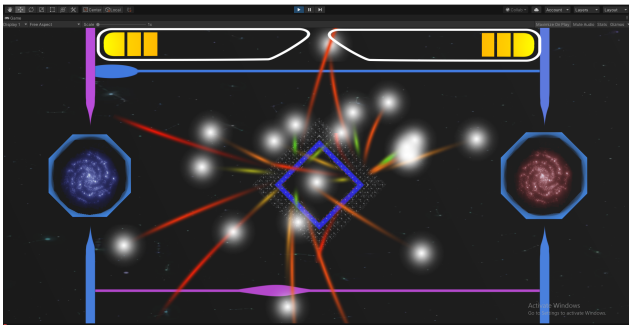


Fig. 16. Graphical Representation of Particle Generation after Collision.

to be invisible for better visual experience as displayed in Fig. 17. A flow chart of proposed algorithm is shown in Fig. 11.

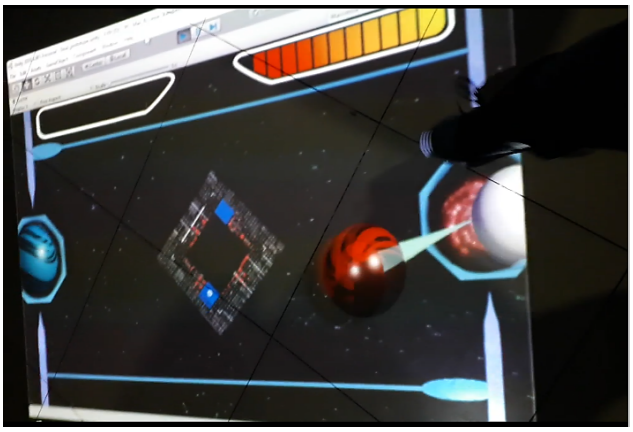


Fig. 17. Practical Testing of Proton-Proton Collision System.

B. Higgs Field

Higgs Field is a field of energy that exists everywhere in the entire universe [34], [35]. The particle known as Higgs boson attracts other particles towards itself to gain mass [36]–[38]. Work on the Higgs field and Higgs boson started in 1964 [39], [40]. The theory of particle was first introduced by PW Higgs in 1964, describing the existence of a particle having a mass of 125 GeV (giga-electron volt) [40]. On 4th of July, 2012, by collaboration of Compact Muon Solenoid and Atlas, the Higgs Boson was first discovered using the LHC tunnel [34], [37].

The proposed system enables to visualize the human body in the presence and absence of the Higgs Field. Thus, explaining the concept of Higgs field and Higgs Boson. The simulated environment shows that when user stands inside the Higgs field, Higgs Boson particles surrounds the user. On a contrast, when body is outside the Higgs field, the particles do not interact with the user. This allows users to visualize that how Higgs Boson interacts with other atoms.

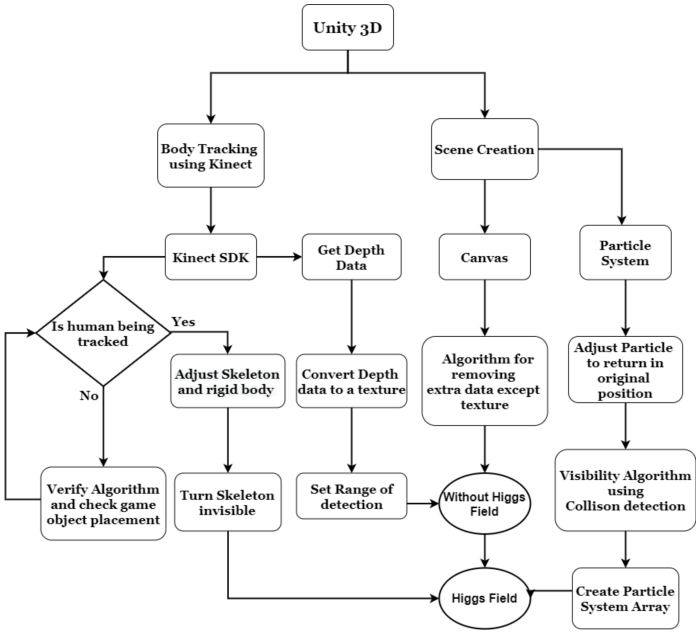


Fig. 18. Higgs Field Flow Chart

1) Implementation of Higgs Field: Firstly, we created an empty 3D project inside Unit 3D, allowing us to develop a system in 3D environment. The same procedure was used to track and create human joints, as in the particle behavior system and it was also made invisible. The script was created to use the depth camera. Data from the depth camera is stored in ushort datatype which cannot be visualized. So, for visualization, we convert depth data from ushort to Texture2D format. A limiter is set which allowed only for the detection of object inside specified z-axis.

A scene was created in which the system was developed. An algorithm was applied to the canvas which allowed it to only make Texture2D datatype visible. The Texture2D enabled us to visualize the behavior of Higgs Boson with other particles in the absence of Higgs field as shown in Fig. 19.

Particle System is created and adjusted to allow the particles to start from outside and return to their respective origins. A script was created and applied to particle system that helps users' to visualize the effect of particle system whenever any object collides with it. An array of particle systems was created using the adjusted particle system and was aligned with the canvas as shown in Fig. 20. The implementation of particle system allowed us to visualize the behavior of Higgs Boson with other particles in the presence of Higgs field. Fig. 18 shows the flow chart of proposed algorithm.

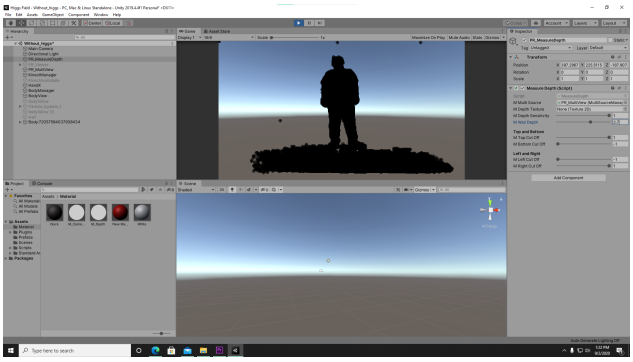


Fig. 19. Depth Camera Data (ushort datatype) converted to Texture2D and implemented on Canvas.

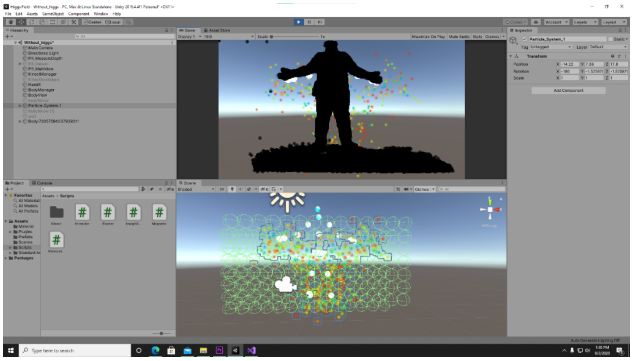


Fig. 20. Visualizing Higgs Field (High Resolution of Canvas).

V. Results and Discussion

This section presents the qualitative and quantitative evaluation of the proposed system. Also the qualitative analysis of the latest AR based learning systems is presented.

A. Qualitative Analysis

The qualitative analysis is based on cost, product originality, scalability, flexibility and market demand as shown in Table III. These are few factors for a product to gain a place in the market and customer attention. The cost factor shows the cost per module of the system. Product originality shows that whether the system developed by the authors is a new product or a revamp (improved version of a previous product). Scalability shows us whether the system can be integrated and scaled towards other fields. Flexibility tells us whether the product or system created can be upgraded later in the future or not. Market Demand is a biased factor which solely depends upon the region of sale. We have considered the sale or outreach of these products on the basis of market trends in Pakistan. The qualitative results suggest that the proposed learning system is highly comparable with recently developed AR based learning systems.

1) Estimated Cost: The proposed system cost around 631 USD. The cost of Projector, Kinect Sensor, and the developed software are the main components. Table II shows the cost of the proposed system.

TABLE II. Cost of Proposed System.

S.No	Items/materials required	Cost in USD
1	Kinect sensor	156
2	Projector	437
3	Software development and installation	38
	Total cost	631

B. Quantitative Analysis

The quantitative analysis is based on a survey as suggested in [21], [29]. A questionnaire was developed for analyzing the effectiveness of the proposed system. Questionnaire shown in Table IV is similar to the questionnaires developed in [21], [29]. In [21], [29], a post-study questionnaire involving scale-based and free-response questions was taken from fifteen participants aged between 21-31 years. Whereas in our study, a total of 20 college students aged between 19 and 22 years participated.

The survey was conducted following the similar pattern as discussed in literature [2], [21], [24], [27], [29]. Firstly, the participants were briefed about the proton-proton collision effect and Higgs field through conventional method of teaching. Then, the proposed AR-based system was used to demonstrate the same concepts. The proposed system allowed the students to interact with the environment and visualize its results. The understanding was made easy, fun-to-learn and interactable.

The survey was performed in a group of 5 students each lasting 20 minutes.

The results attached in Fig. 21 were analyzed on the basis of the methods used in [21], [29]. It mainly focuses on the effectiveness of the AR based learning systems in learning process.Whereas, Fig. 22 shows the score of each questions. Similarly, the result of first four questions focuses on the improvements in the understanding of topics. Whereas, the Q5 result shows the response of students about recommendation of proposed system as shown in Fig. 23. It was concluded that 85% of the students recommended the proposed system for effective learning.

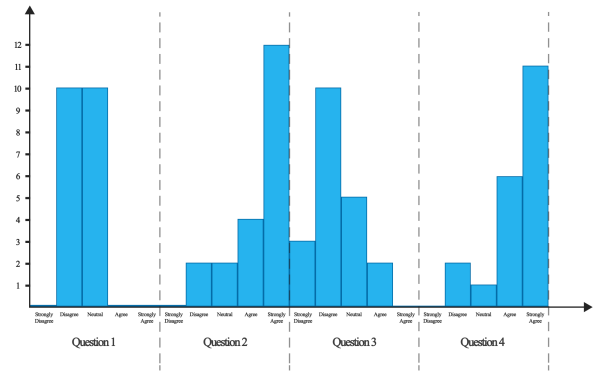


Fig. 21. Survey Results.

VI. Limitations of the Proposed System

Following are the limitations of the proposed system:
1. The total horizontal and vertical field view of Kinect sensor

TABLE III. QUALITATIVE ANALYSIS OF RECENT AR BASED LEARNING SYSTEMS AND PROPOSED SYSTEM.

Source	Cost in USD	Product Originality	Scalability	Flexibility (future work)	Market Demand (Value)
[19]	5k	New product	High	High	Medium-high
[12]	950	Revamp	High	High	Medium-high
[3]	350	Revamp	Medium	High	Low-Medium
[21]	650	Revamp	Medium	Medium	Low-to-medium
[22]	1.3K	New product	High	High	Medium-High
[23]	625	Revamp	Medium	Medium	Low-to-medium
[24]	625	Revamp	Low	Medium-Low	Low
[25]	625	Revamp	Medium	Medium-High	Medium-high
[27]	375	Revamp	Medium	High	Medium-High
[28]	375	Revamp	Medium	High	Low-Medium
[29]	63	New product	Medium	Medium-High	Medium-High
Proposed System	631	New product	High	High	Medium - High

TABLE IV. PROPOSED SYSTEM SURVEY (5 MULTIPLE-CHOICE QUESTIONS).

Number	Question Details
Q1	It is easy to understand proton-proton collision effect without AR-based demonstration
Q2	I found AR-based demonstration very helpful in understanding the proton-proton collision experiment
Q3	It is easy to understand the concept of higgs field without AR-based demonstration
Q4	I found AR-based demonstration very helpful in understanding higgs field experiment
Q5	I highly recommend this product in education which can help student to understand complex problem

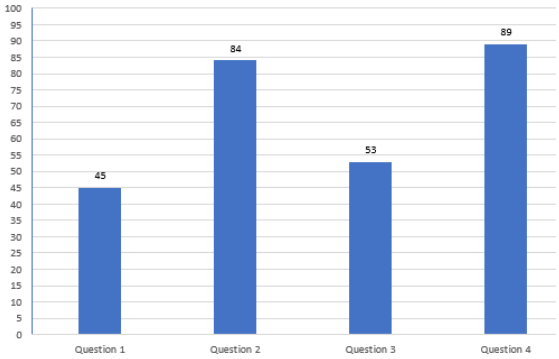


Fig. 22. Results of Individual Questions.

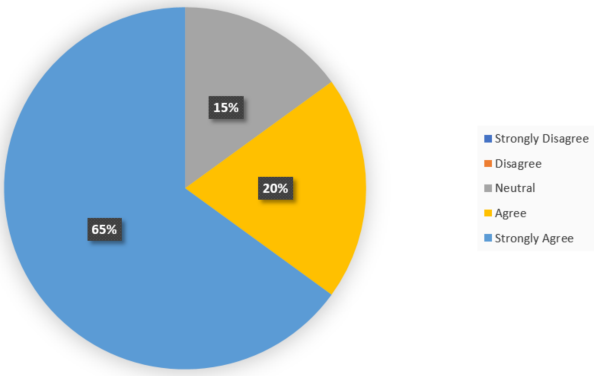


Fig. 23. Results of Project Recommendation Question (Q5).

is 70 and 60 degrees respectively. Hence the proposed system is subjected to a limited space. In order to cover entire room (360 degrees), more Kinect sensors are needed to be linked together.

2. The proposed system can smoothly detect up to two people at a time. Hence putting a constraint on certain application where more than two people needs to be detected at the same time. Using more Kinect sensors can also solve this issue.

3. If the user performs a fast movement, there is an issue of delay. Hence depth camera requires some extra time to detect each movement correctly. This issue can be resolved using Graphical Processing Unit with high memory specifications.

4. The Kinect sensor has to be fixed and aligned with the position of users for smooth experience of users.

5. The Kinect sensor has a specific detection range of around 10 meters.

6. The proposed system works better in a room where there is minimal or no sunlight.

VII. CONCLUSION

In this research paper, the detailed review on latest advancement in AR-based learning systems is presented. The detailed review is based on major contributions and limitations provided by authors. It is concluded through intensive review that AR-based learning systems are effective to incorporate 21^{st} century skills such as critical thinking, creativity, communication, collaboration, innovation and problem solving. Furthermore, an AR-based learning system is developed to demonstrate the particle physics experiments. The proposed system simulates the proton-proton collision and Higgs field

using unity 3D software. The Proton-Proton collision algorithm simulates the generation of new particles when protons collide with each other. It demonstrates that number of particles generated depends on energy of the colliding protons. Whereas, Higgs field simulation highlights the effect of Higgs Boson inside and outside Higgs field. The proposed system is developed using unity 3D software and then interfaced to Kinect sensor for immersive experience. Then, the qualitative analysis of the proposed system and latest AR based learning systems is presented. Finally, the quantitative analysis of the proposed system is conducted. Overall, the results suggest that 85% of the participants recommended the proposed learning system.

REFERENCES

[1] J. Egger and T. Masood, "Augmented reality in support of intelligent manufacturing – A systematic literature review," Comput. Ind. Eng., vol. 140, p. 106195, 2020, doi: 10.1016/j.cie.2019.106195.

[2] P. Fraga-Lamas, T. M. Fernández-Caramés, Ó. Blanco-Novoa, and M. A. Vilar-Montesinos, "A Review on Industrial Augmented Reality Systems for the Industry 4.0 Shipyard," IEEE Access, vol. 6, pp. 13358–13375, 2018, doi: 10.1109/ACCESS.2018.2808326.

[3] L. Haz, Y. Molineros, E. Vargas, and A. Davila, "Multimedia system Kinect-based. Learning experience for children of primary school," CEUR Workshop Proc., vol. 2486, pp. 295–308, 2019.

[4] S. B. Adikari, N. C. Ganegoda, R. G. N. Meegama, and I. L. Wanniarachchi, "Applicability of a Single Depth Sensor in Real-Time 3D Clothes Simulation: Augmented Reality Virtual Dressing Room Using Kinect Sensor," Adv. Human-Computer Interact., vol. 2020, 2020, doi: 10.1155/2020/1314598.

[5] G. Dini and M. D. Mura, "Application of Augmented Reality Techniques in Through-life Engineering Services," Procedia CIRP, vol. 38, pp. 14–23, 2015, doi: 10.1016/j.procir.2015.07.044.

[6] Y. Tokuyama, R. P. C. J. Rajapakse, S. Yamabe, K. Konno, and Y. P. Hung, "A kinect-based augmented reality game for lower limb exercise," Proc. - 2019 Int. Conf. Cyberworlds, CW 2019, pp. 399–402, 2019, doi: 10.1109/CW.2019.00077.

[7] Di. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui, "Mobile Augmented Reality Survey: From Where We Are to Where We Go," IEEE Access, vol. 5, pp. 6917–6950, 2017, doi: 10.1109/ACCESS.2017.2698164.

[8] H. Peng, "Application Research of AR Holographic Technology based on Natural Interaction in National Culture," Proc. 2019 IEEE 4th Adv. Inf. Technol. Electron. Autom. Control Conf. IAEAC 2019, no. Iaeac, pp. 2220–2224, 2019, doi: 10.1109/IAEAC47372.2019.8997672.

[9] S. M. C. Loureiro, J. Guerreiro, and F. Ali, "20 years of research on virtual reality and augmented reality in tourism context: A text-mining approach," Tour. Manag., vol. 77, no. October 2019, 2020, doi: 10.1016/j.tourman.2019.104028.

[10] A. Annafi, D. L. Hakim, and D. Rohendi, "Impact of using augmented reality applications in the educational environment," J. Phys. Conf. Ser., vol. 1375, no. 1, 2019, doi: 10.1088/1742-6596/1375/1/012080.

[11] A. Moore et al., "Comparative usability of an augmented reality sandtable and 3D GIS for education," Int. J. Geogr. Inf. Sci., vol. 34, no. 2, pp. 229–250, 2020, doi: 10.1080/13658816.2019.1656810.

[12] P. D. Petrov and T. V. Atanasova, "The Effect of augmented reality on students' learning performance in stem education," Inf., vol. 11, no. 4, 2020, doi: 10.3390/INFO11040209.

[13] S. Cai, X. Wang, and F. K. Chiang, "A case study of Augmented Reality simulation system application in a chemistry course," Comput. Human Behav., vol. 37, pp. 31–40, 2014, doi: 10.1016/j.chb.2014.04.018.

[14] M. Kesim and Y. Ozarslan, "Augmented Reality in Education: Current Technologies and the Potential for Education," Procedia - Soc. Behav. Sci., vol. 47, no. 222, pp. 297–302, 2012, doi: 10.1016/j.sbspro.2012.06.654.

[15] J. Porozovs and S. Kristapsone, "The Opinion of Latvian Teachers About the Most Suitable Teaching Methods and Possibilities to Make Lessons Interesting," J. Pedagog. Psychol. "Signum Temporis," vol. 9, no. 1, pp. 50–56, 2019, doi: 10.1515/sigtem-2017-0009.

[16] N. Pellas, P. Fotaris, I. Kazanidis, and D. Wells, "Augmenting the learning experience in primary and secondary school education: a systematic review of recent trends in augmented reality game-based learning," Virtual Real., vol. 23, no. 4, pp. 329–346, 2019, doi: 10.1007/s10055-018-0347-2.

[17] P. Fraga-Lamas, T. M. Fernández-Caramés, Ó. Blanco-Novoa, and M. A. Vilar-Montesinos, "A Review on Industrial Augmented Reality Systems for the Industry 4.0 Shipyard," IEEE Access, vol. 6, pp. 13358–13375, 2018, doi: 10.1109/ACCESS.2018.2808326.

[18] S. R. Dehkordi, M. Ismail, and N. M. Diah, "A review of kinect computing research in education and rehabilitation," Int. J. Eng. Technol., vol. 7, no. 3, pp. 19–23, 2018, doi: 10.14419/ijet.v7i3.15.17399.

[19] F. Garzotto, M. Gelsomini, M. Gianotti, and F. Riccardi, "Engaging children with neurodevelopmental disorder through multisensory interactive experiences in a smart space," Internet of Things, vol. 0, pp. 167–184, 2019, doi: 10.1007/978-3-319-94659-7_9.

[20] M. S. D. R. Guerra and J. Martin-Gutierrez, "Evaluation of full-body gestures performed by individuals with down syndrome: Proposal for designing user interfaces for all based on kinect sensor," Sensors (Switzerland), vol. 20, no. 14, pp. 1–22, 2020, doi: 10.3390/s20143930.

[21] N. J. Sung, J. Ma, Y. J. Choi, and M. Hong, "Real-time augmented reality physics simulator for education," Appl. Sci., vol. 9, no. 19, pp. 1–12, 2019, doi: 10.3390/app9194019.

[22] T. Matuszka, F. Czuczor, and Z. Sóstai, "Heromirror interactive: A gesture controlled augmented reality gaming experience," ACM SIGGRAPH 2019 Posters, SIGGRAPH 2019, pp. 2–3, 2019, doi: 10.1145/3306214.3338554.

[23] R. Lozada-Yánez, N. La-Serna-Palomino, and F. Molina-Granja, "Augmented Reality and MS-Kinect in the Learning of Basic Mathematics: KARMLS Case," Int. Educ. Stud., vol. 12, no. 9, p. 54, 2019, doi: 10.5539/ies.v12n9p54.

[24] C. Pittman and J. J. L. V. Jr, "PhyAR: Determining the Utility of Augmented Reality for Physics Education in the Classroom," Proc. - 2020 IEEE Conf. Virtual Real. 3D User Interfaces, VRW 2020, pp. 761–762, 2020, doi: 10.1109/VRW50115.2020.00231.

[25] M. Zhang, Z. Zhang, Y. Chang, E. S. Aziz, S. Esche, and C. Chassapis, "Recent developments in game-based virtual reality educational laboratories using the microsoft kinect," Int. J. Emerg. Technol. Learn., vol. 13, no. 1, pp. 138–159, 2018, doi: 10.3991/ijet.v13i01.7773.

[26] D. Alexandrovsky, S. Putze, T. Stabbert, T. DÖring, T. FrÖhlich, and R. Malaka, "Demonstrating Vrbox-a virtual reality augmented sandbox," Conf. Hum. Factors Comput. Syst. - Proc., pp. 1–4, 2019, doi: 10.1145/3290607.3313251.

[27] M. Xu et al., "Personalized training through Kinect-based games for physical education," J. Vis. Commun. Image Represent., vol. 62, pp. 394–401, 2019, doi: 10.1016/j.jvcir.2019.05.007.

[28] Y. H. Chang, P. R. Lin, and Y. Te Lu, "Development of a kinect-based english learning system based on integrating the ARCS model with situated learning," Sustain., vol. 12, no. 5, 2020, doi: 10.3390/su12052037.

[29] M. C. Costa, A. Manso, and J. Patrício, "Design of a mobile augmented reality platform with game-based learning purposes," Inf., vol. 11, no. 3, pp. 1–20, 2020, doi: 10.3390/info11030127.

[30] K. E. Uhm et al., "Usefulness of Kinect sensor–based reachable workspace system for assessing upper extremity dysfunction in breast cancer patients," Support. Care Cancer, vol. 28, no. 2, pp. 779–786, 2020, doi: 10.1007/s00520-019-04874-2.

[31] A. Anwer, S. S. Azhar Ali, A. Khan, and F. Meriaudeau, "Underwater 3-D Scene Reconstruction Using Kinect v2 Based on Physical Models for Refraction and Time of Flight Correction," IEEE Access, vol. 5, pp. 15960–15970, 2017, doi: 10.1109/ACCESS.2017.2733003.

[32] Aaboud M, Aad G, Abbott B, Abdinov O, Abeloos B, Abhayasinghe DK, et al. ATLAS Collaboration. Nucl Phys A. 2019;982:985–1009.

[33] "Particle production in proton-proton collisions M. T. Ghoneim," pp. 1–13.

[34] B. Horn, "The Higgs Field and Early Universe Cosmology: A (Brief) Review," Physics (College. Park. Md)., vol. 2, no. 3, pp. 503–520, 2020, doi: 10.3390/physics2030028.

[35] Higgs field - Simple English Wikipedia, the free encyclopedia. https://simple.wikipedia.org/wiki/Higgs-field#cite-ref-eb64-1-0. Accessed 4 Sep 2020

[36] G. Aad et al., "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," Phys. Lett. Sect. B Nucl. Elem. Part. High-Energy Phys., vol. 716, no. 1, pp. 1–29, 2012, doi: 10.1016/j.physletb.2012.08.020.

[37] "A new boson with a mass of 125 GeV observed with the CMS experiment at the Large Hadron Collider (Science (2012) (1569))," Science, vol. 339, no. 6125. p. 1275, 2013, doi: 10.1126/science.339.6125.1275-b.

[38] P. Azzi, "First look at the physics case of TLEP," Nuovo Cimento della Societa Italiana di Fisica C, vol. 37, no. 2. pp. 11–18, 2014, doi: 10.1393/ncc/i2014-11730-6.

[39] "Broken Symmetries and the Masses of Gauge Bosons - PhysRevLett.13.508." [Online]. Available: http://journals.aps.org/prl/pdf/10.1103/PhysRevLett.13.508.

[40] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons," Phys. Rev. Lett., vol. 13, no. 9, pp. 321–323, 1964, doi: 10.1103/PhysRevLett.13.321.

Parallelization Technique using Hybrid Programming Model

Abdullah Algarni[1], Abdulraheem Alofi[2], Fathy Eassa[3]
Department of Computer Science, King Abdulaziz University (KAU)
P.O. Box 80221, Jeddah 21589, Saudi Arabia

***Abstract*—A multi-core processor is an integrated circuit that contains multiple core processing unit. For more than two decades, the single-core processors dominated the computing environment. The continuous development of hardware and processors led to the emergence of high-performance computers that able to address complex scientific and engineering programs quickly. Besides, running the software codes sequentially increases the execution time in huge and complex programs. The serial code is converted to parallel code to improve the program performances and reduce the execution time. Therefore, parallelization helps programmers solve computing problems efficiently. This study introduced a novel automatic translation tool that converts serial C++ code into a hybrid parallel code. The study analyzed the performance of the proposed S2PMOACC tool using linear algebraic dense matrix multiplication benchmarking. Besides, we introduced Message Passing Interface (MPI) + Open Accelerator (OpenACC) as a hybrid programming model without preliminary knowledge of parallel programming models and dependency analysis of their source code. The research outcomes enhance the program performances and decrease the implementation time. Moreover, our proposed technique offers better performance than other tools.**

***Keywords*—Serial code translation; parallel code; C++; hybrid programming model; auto-translation; S2PMOACC**

I. Introduction

A single-core microprocessor dominated the computing environment for more than two decades as it offered better performance in execution of computer programs. With a rise in issues, such as power dissipation, design complexity, and high energy consumption [1] in the single-core, multicore architectures were proposed to address these problems. The multicore architectures [2] opened a new door for high-performance computing, dividing each task into different cores during execution. Also, the multi-core architecture plays a crucial role in developing parallel applications. Therefore, many industries build their programs using parallel computing architectures [3]. Besides, complex scientific programs require a huge computing power [4], which individual computer fails to provide. Therefore, programmers must write parallel programs to be running in multicore architectures. Also, parallel programming is computationally complex and requires different execution effort.

Nowadays, with the advent of computer technology, people rely heavily on computer systems to conduct all business-related tasks. A standard desktop computer or workstation can easily solve small computing problems, but it provides poor performance and runs into technical problems while performing a high number of operations per second. Besides, a standard computer faces challenges in completing a time-consuming operation in less time, completing operations under a tight deadline, and solving complex large problems. High-Performance Computing (HPC) eliminates these problems. Compared to a traditional desktop computer, HPC systems can perform complex calculations and process data at high speed. Fig. 1 shows HPC systems with CPUs. Each of the CPUs processors has local memory and multicores, which help to execute different applications and challenging tasks. Besides, HPC solves extensive problems via thousands of parallel processors.

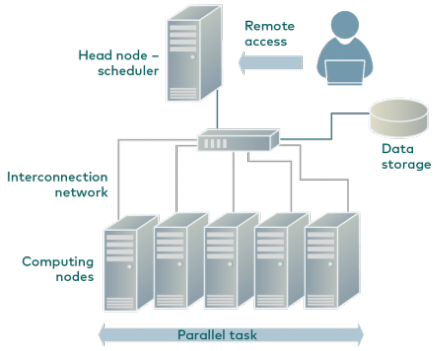


Fig. 1. An Overview of HPC System.

The HPC offers significant benefits to different sectors such as education, health, engineering, government, and business owing to its ability to solve complex and demanding problems. Additionally, HPC is an essential driver of innovation and fosters economic growth. Besides, HPC facilitates R&D in science and technology and enhances new products time-to-market. HPC also helps researchers to solve complex problems, such as developing new drugs through advanced computer simulation. Parallel computing is a computation that breaks larger computing problems into smaller tasks, in which many calculations are executed concurrently. Massive problems are divided into smaller units in order to enhance the overall performance of HPC systems. Besides, the development of future Exascale machines can become complex, which requires writing parallel programs [5], [6]. Parallel programming [3] is a multi-threaded or multi-processes mechanism used to write and run the parallel programs on the HPC. There are many existing parallel programming languages like OpenMP (Open Multi-Processing), OpenACC (Open accelerators) and MPI (Message Passing Interface).

Creating parallel programs manually is a hard job and leads to consuming time and so may not free from human mistakes [6]. Thus [7], programmers are increasingly using automatic parallelization tools owing to their ability to automatically translate serial code into parallel code, thereby saving programming time and costs. Many automatic parallelization tools we will discuss in the related works section, these tools are available to convert sequential codes to parallel codes in order to minimize programming errors, and offer accurate results [7]. Besides, these tools take different inputs and combine different programming models. Combining more than one model is crucial in achieving optimal performance through parallelizing programs [8]. Therefore, the hybrid MPI+OpenMP [5] model is perfect for parallel computing,because it is combine between shared and distributed memory. Clearly, each tool only suitable for specific parallel model [9], and no tool is good enough for all applications .

The current study highlighted there is no parallelization tool exists on the cluster/hybrid system that converts the serial code into parallel. In order to leading towards the objective, we develop a hybrid MPI-OpenACC tool able to translate sequence C++ codes to parallel codes, implemented by combining the MPI library with OpenACC directives. Fig. 2 demonstrates how the hybrid model works [10]. The hybridization increases performance [11], parallelism, and adapts to different environments.

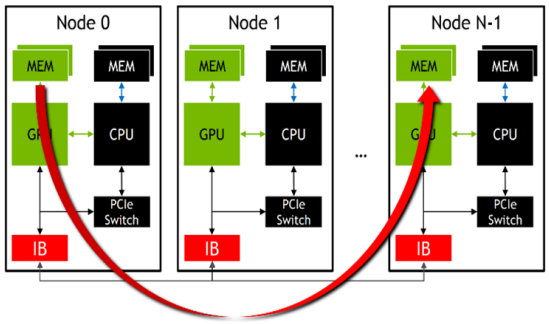


Fig. 2. Processing Mechanism of Hybrid MPI and OpenACC.

The study provides the following contributions: The study proposes a new automatic translation tool that converts serial C++ programming code into hybrid MPI+OpenACC parallel code. Besides, we propose an algorithm and theoretical architecture to enhance performance and decrease implementation run-time. We implement many applications using the proposed solution and compute the results on the different Graphics Processing Units (GPU) devices. Furthermore, the performance and features of the proposed model are compared with existing automatic tools. Based on experimental results, our proposed technique outperforms other models.

The rest of the paper is organized as follows. In Section 2, we discuss the detailed background of parallel computing models used in our proposed solution and then, Section 3 describes the related works of parallel programming models based on different hierarchical machines. Section 4, the system model has been described in detail with the architecture, and algorithm of the proposed parallelization technique. Section 5, discusses the experimental platform and the measuring factors for evaluating the proposed technique. Section 6, provides results and discussion. Finally, the conclusion in Section 7 followed by future work in Section 8.

II. Background

The rapid development of hardware and processors led to the emergence of parallel computing, which can address complex scientific and engineering programs quickly and efficiently. In parallel computing, the program is broken into several parts to solve computation problems concurrently. Each part is further broken into a set of instructions to be executed simultaneously on different processors. The primary benefit of parallel computing is suitability for modeling and simulation. Besides, parallel computing saves time and produces useful results for researchers. Single-core processors are unsuitable to solve many scientific problems. However, multi-core processors can solve these problems as they contain GPU. As shown in Fig. 3, the MPI parallel programming model standard was launched in 1994 for the application of distributed-memory communication.

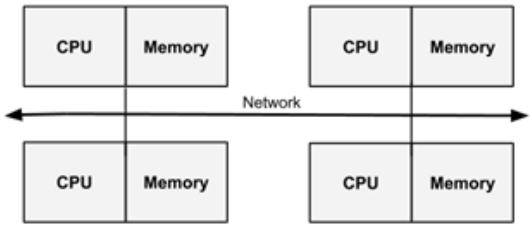


Fig. 3. MPI Distributed Memory.

MPI [12] is an efficient standard programming model applicable on distributed computing systems for many years, MPI is implemented on parallel machines and provides good performance and portability. MPI implementations are designed to work on different parallel environments and support classical communications [13]. MPI provides an explicit method for the message passing a programming technique on distributed memory clusters. Besides, distributed resources are spread over several computing nodes, and in MPI, synchronization is handled explicitly due to the distributed memory [14]. Many attributes in MPI such as portability, where it is has a ability to integrate with other programming models. Also, it is available for many implementations, such as open-source implementations like MPICH [15], and OpenMPI [16], and commercial implementations such as Intel MPI library [17], and IBM Spectrum MPI [18]. Further, functionality [11], where the MPI library has more than 400 routines. MPI programs have special structures and listings. First, demonstrating the basic commands, starting from the MPI header file. Then, initialize the MPI environment using MPI_Init() instruction. The next stage is by defining the rank and the size of processes using MPI_Comm_rank and MPI_Comm_size consecutively. In another stage, inserting MPI calling routine code and run parallelly. Finally using MPI_Finalize() to terminate the MPI execution [19]. With the advancement in GPU technologies, accelerators have been developed for GPU programming. Each accelerator follows a unique programming technique. For example, Compute Unified Device Architecture (CUDA) for

NVIDIA GPUs, Brook+ for AMD GPUs, and LEO for MICs, etc. [20]. NVIDIA released CUDA in 2007. In November 2011, OpenACC introduced as directive-based programming model designed for targeting heterogeneous CPU/GPU systems. OpenACC [11] has features to overcome the limitation of the CUDA model. CUDA works on NVIDIA GPU only whereas OpenACC works with many compilers. Besides, OpenACC offers excellent performance and accelerates scientific applications with little programming efforts [21]. The programmer only should insert directives in a suitable place to run the code on the GPU compiler [7]. Fig. 4 shows the OpenACC accelerator model, revealing how intensive computations are offloading from a host device to the GPU device to accelerate the execution [22].

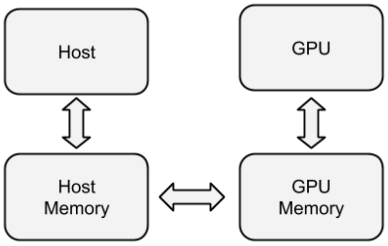


Fig. 4. OpenACC Accelerator.

Accelerating the program through OpenACC requires analyzing the source code to determine the part that requires more execution time. As illustrated in Listing 1, inserting the directives in a suitable place facilitates the code execution on the GPU compiler.

Listing 1: OpenACC directives

```
main()
{
  <serial code>
  #pragma acc kernels
  //automaticly runs on GPU
  {
        <parallel code>
  }
}
```

III. Related Works

In this section, we have discussed the detailed hybrid programming models and auto parallelization tools to convert serial code to parallel. The programming model combined with one or more models can increase the performance of parallelism and the capability to work with the heterogeneous systems. This combination [23] facilitates the application of large-scale powerful programming models. Fig. 5 depicts the hierarchy navigation programming model and is categorized as follows:

- Single model: MPI
- Dual model: MPI+X
- Tri model: MPI+X+Y

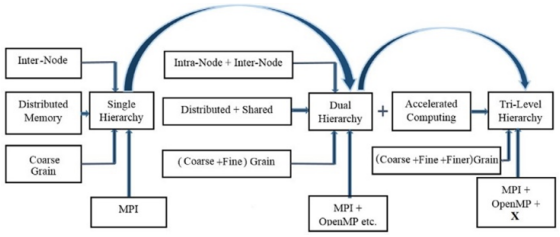


Fig. 5. Hierarchy Navigation in the Programming Model.

1) Single model: MPI
MPI is a standard message passing library for exchanging data between different nodes. MPI facilitates program execution among distributed memory and is an effective mechanism to parallelize the application [12]. Besides, it is a coarse-grained technique to execute and manage data on the level of the node [24]. The MPI version 3.1 introduces new features and capabilities to facilitate the parallelization process like creating group quires and processes [13], [12]. The HPC environments help to share data across different distributed nodes using the MPI library. Therefore, MPI programmers must understand future hardware development for effective compatibility as MPI provides an excellent model for future disparate systems.

2) Dual model: MPI+OpenMP
Previous studies [25], [26] introduced a common hybrid model MPI + OpenMP, using MPI for communication between nodes. Besides, OpenMP is used for the parallelization process inside the node. Fig. 6 shows the processing mechanism of hybrid MPI and OpenMP. Data are shared over several nodes that communicate with each other through the MPI message passing technique. The OpenMP region is designed for distributed data, assisting in deciding the available number of threads. This hybrid approach is one of the promising models for future HPC applications [27].

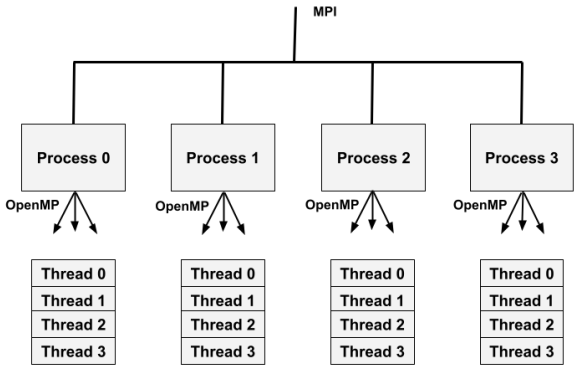


Fig. 6. Hybrid MPI and OpenMP Processing Mechanism.

3) Dual model: MPI+OpenACC

To use parallel programming, Hybrid MPI and OpenACC are an alternative model for hybrid MPI and CUDA mentioned in [28]. Similar to OpenMP, data parallelization in OpenACC is based on directives. The programmers must write their code, interchangeable to traditional serial code written in C/C++, and include the #pragma acc loop directive line before the loop statements. The program written in these directives is computed with accelerated GPU devices. In theory, the mixed mode architecture code should be more efficient more than a pure MPI model due to combination between shared and distributed memory. Listing 2 describes a briefing algorithm as a hybrid of MPI + OpenACC is the simplest way of parallelism [27].

Listing 2: MPI+OpenACC Processing

```
MPI_Init() //Initialize MPI
//Processes size
Size <-- Get MPI_Comm_size()
//Processes ranks
Rank <-- Get MPI_Comm_rank()
if (Rank==0) //Master Process
/* Make processing
 before Entering MPI comm world */
//Send data when rank > 0
MPI_Isend()
//Check processing periodically
MPI_Wait()
if(Rank>0)
/* receive data from
all processes rank > 0 */
MPI_Irecv()
 #pragma acc data copy(a) copyin(b)
 #pragma acc kernels
 {
   While(loop_statements 1 to N)
   #pragma acc loop
   loop statement 1
 }
MPI_Isend() //send data again
if(Rank==0) //collect data
MPI_Irecv() //receive final data
MPI_finalize() //finalize MPI
```

4) Tri model: MPI + OpenMP + CUDA (MOC)
In 2018, a group of developers introduced the Tri-Hierarchy hybrid MOC model [23], comprising of MPI + OpenMP + CUDA to achieve enormous parallelism objectives. MPI helps to broadcast data on the distributed node. OpenMP executes data on CPU threads; whereas, CUDA executes data on accelerated GPU cores [29]. Besides, the MOC model facilitates performance via inflexible parallelism. We develop a similar HPC application using a huge cluster system with multiple nodes and GPU. MOC model offers a coarse and efficient massive parallelism.

As discussed previously, writing parallel code manually is tedious and time-consuming. The auto parallelization tools help to overcome these problems. Many tools can convert sequential codes to parallel codes and add the parallelization constructs or directives in a suitable place [9].

One study [30] provides a concise survey of existing parallel tools and classifies these tools based on different criteria like a history of tools, tools contributions, and support assisting for parallelization. A new tool called **EasyPar** proposed in the study and this tool capable to assisting and facilitating the program's development phase. The authors in [31] analyze the performance of two tool(Cetus and Par4All). The **Cetus** is a source-to-source transformation tool using OpenMP directives to convert 'C' sequential code to parallel codes. However, the **Par4All** is an open-source tool to convert sequential code to new OpenMP, CUDA, and OpenCL source codes. The study performed on two complex program [31] to find which the best performance between tools. The results showed the tools are effective for single loop programs but not for the nested loops. A study [32] proposes a new parallelization model called **PyParallelize**, which automates the parallel process by reading source code without modifications from the programmer. After implementing the model on different benchmark programs, the results showed a relatively high rate of accuracy. However, this model fails to work efficiently when nested loops are more than two in source code. Another study [4] presents a new converting tool called **S2P** (c serial to Parallel) to perform parallelization on different programs. Furthermore, they [4] compare the tool performance with other existing tools. The **S2P** tool offers better results in some cases and it assists in minimizing the overhead thread management during the execution of parallel code. Authors in [9] proposed an automatic code parallelization tool, which converted C serial code to the equivalent version of parallel using OpenMP parallel programming. However, this tool focuses only on parallelism tasks without considering loops parallelism.

A new model architecture suggested in the study [33], and that model can translate any serial application into parallel code, using individual parallel programming likes OpenMP, MPI, OpenCL, and OpenACC or hybrid OpenMP-MPI. This model is under development with a promise to solve automatic parallelization issues. Besides, a study [7] proposed a tool to speed up multicore processors. The proposed tool achieves 4.27 speedup after using 4 cores and 8 threads when increasing the length of the matrix.

We conclude from the previous literature, there is no tool provides full optimal translation. Thus, research is required to address the limitations of these tools. Tools such as SUIF, CAPO, and Polaris fail to support all operating systems. Also, these tools only used C and FORTRAN languages [30]. Besides, it is difficult to run the generated code on distributed memory because the tools such as Par4All and S2P use the OpenMP programming model, which works exclusively on shared memory. The parallelization in nested loops are an advantage of the tool. Therefore it is a limitation in the tools like SUIF and Intel compilers [32]. In the next section, we will discuss in detail the methodology, revealing the proposed S2PMOACC tool.

IV. Methodology

This study introduces a new automatic translation tool to reduce the software code's execution time sequentially by massively improving the performance of huge and complex programs. We use all the available resources to convert serial code to parallel code to enhance the program performances and reduce the execution time. We proposed a new automatic translation tool to convert serial C++ code into parallel code. Fig. 7 illustrates the proposed translation tool S2PMOACC (Serial To Parallel MPI and OpenACC) architecture, taking serial code as input and generating its parallel code automatically. The proposed solution enhances program performance and decreases implementation time.

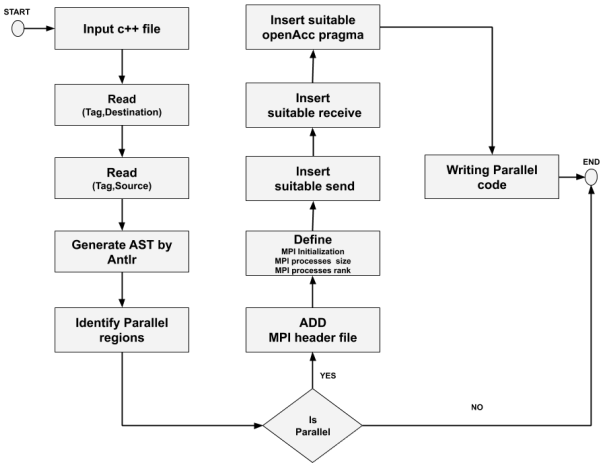


Fig. 7. Architecture of Proposed Automated Translation Tool S2PMOACC.

The steps of the translation tool are elaborated as follows:

1) Input C++ file

   This is the first phase where the source file enters the command line using the CLC library. The C++ source file must be written without any syntax error. Then, the file name is moved to the next steps for further processing.

2) Read (tag.Destination)

   This step reads input from the user and enters the information related to the MPI_Send call. Using tag in the MPI to distinguish between different processes and the destination is an INT number for process rank where every process has a unique rank in the MPI environment. This information should be provided by the user to write statements when generating the parallel code.

3) Read (tag.Source)

   As the previous step, the user enters information of tag and source related to MPI_Recv call. Source is an INT number to determines the rank of the source process. To write a suitable receive statement, the information must be entered correctly.

4) Generate AST by ANTLR

   ANTLR tool will read the file contents and generate the AST to be used in the following steps. As shown in Fig. 8, many procedures in ANTLR are revealed. First, reading a file's contents by passing the name to a specific routine to handle the process. Then, the contents streaming fed to lexer that contains lexer grammars to identify and produce different token streams. Finally, the token streams are entered into the parser to generate AST based on pre-defined parser rules.

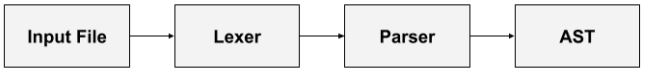


Fig. 8. ANTLR Tool Phases.

5) Identify parallel regions

   In this step, the generated AST is used to identify the parallel regions of the code using inherited class to access and identify different loops and variables.

6) Determine the dependency

   In this step, the dependency test occur to determine the possibility of parallelization. Special java class used to find out parallel regions like loops. The dependency test used to: accessing loop, fetch statements in the loop body, and decide if dependency inside the body block existing or not. For example if **S1** and **S2** are statements inside the loop, if **S2** depend on **S1** result, then the dependency is detected. Therefore, the target file is generated without parallel directives. Otherwise, the dependency not found and parallel file generated.

7) Add MPI header file

   Here, include MPI header file library 'mpi.h' in order to use MPI routines when generating parallel code.

8) Define MPI instructions

   Initialize MPI environment through using MPI_Init() instruction. Then define Ranks for MPI to check cluster system specification for defining the ranks via MPI_Comm_rank that provide a logical way of numbering the MPI process. Further, define MPI communication world statements via MPI_Comm_world in order to run all MPI jobs across several processes. Ranks in this step used to find out master and slave processes in MPI environment.

9) Insert suitable MPI_ Send statement

   In order to write a suitable MPI send statement in the parallel file the information get it from a user in the previous steps will be used. The statement will be insert based on a pre-defined indicator from the user. Also, the information should be correct to avoid mistakes in writing process.

10) Insert suitable MPI_ recv statement

The information entered from the user is used in this step. Besides, the pre-defined indicator is used to identify the place to insert receive MPI operation.

11) Insert suitable OpenACC directive

Insert OpenACC directives based on the parallelization annotations. This step involves inserting an OpenACC directive to the source code to notify the OpenACC-compiler to parallelize the objects of the classes, run in parallel during run-time.

12) Writing Parallel code

Write parallel code to generate the code, which includes the MPI routines and OpenACC parallel computing pragmas via #pragma acc atomic and #pragma acc parallel loop.

Regarding the proposed automated translation, we discuss a comprehensive overview of how C++, MPI, and OpenACC are translated from serial code to dual auto parallelization using a hybrid programming model. Algorithm 1 presents serial code to dual auto parallelization.

Algorithm 1 : Serial Code to Dual Auto Parallelization

1: Input C++ file.
2: Read (Tag, Destination) information.
3: Read (Tag, Source) information.
4: Generate AST by ANTLR.
5: Identify parallel regions(for-while-do while)
6: Perform dependency analysis for each region.
7: Determine the possibility of parallelization.
8: Add MPI header file 'mpi.h'.
9: Initialize MPI environments 'MPI_Init()'
10: Define MPI communication world statements
11: Define MPI communications ranks
12: Insert suitable MPI_Send().
13: Insert suitable MPI_Recv().
14: Insert OpenACC directives based on the parallelization annotations.
15: Writing parallel code.
16: Save the output file in the directory.

The next sections will shows the experimental platform and measuring factors followed by results and discussion for testing the proposed tool and illustrates the capabilities and limitations of our tool compared with other existing tools.

V. Experimental Platform and Measuring Factors

This section demonstrates the experimental platform to analyze the performance of the proposed solution. We quantified different measures, taking HPC benchmarks as performance metrics of the execution time (Secs) and system speedup (Serial/Parallel) to compare the proposed tool and other tools. We perform all the experiments on Intel i7 with four cores and eight threads. Table I shows the testing environment specifications. The HPC system and applications are evaluated based on fundamental performance metrics. We measured different performance attributes including execution time (Secs) and speedup (Serial/Parallel) of the system.

TABLE I. Environment Specifications

Feature	value
Processor Type	Intel(R) Core(TM) i7-1065G7
Number of cores	4
Number of threads	8
Operating Systems	Windows 10
Clock speed	1.30 GHz
Graphic card	NVIDIA GeForce MX250
RAM	16 GB

To evaluate the performance attributes, we select a dense matrix multiplication (DMM) with different sizes and computed on the different numbers of MPI processes and GPUs devices. We evaluate the speed up performance metric where matrix multiplication is computed without using a single number of GPU core to determine the speed trend. Without matrix multiplication, we run speed on the experimental setup and calculate the time taken in sequential processing. Ideally, the speedup is calculated by following the fundamental Amdahl's law [34] and Gustafson [35].

$$SpeedUp(S_P) = \frac{T_{serial}}{T_{parallel}} \quad (1)$$

Where T_{serial} is the optimal time in sequential processing and $T_{parallel}$ for parallel computing algorithms. According to (1), we can calculated speedup based on execution time by implementing proposed solution on DMM benchmarking application against varying dataset. The possible curve of speedup [36] in an algorithm could be super-linear, perfect linear, linear and sub-linear as demonstrate in Fig. 9.

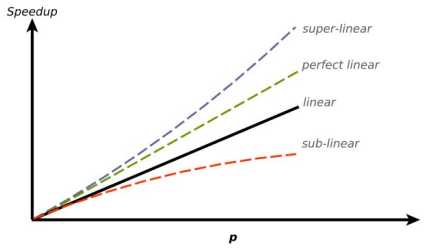


Fig. 9. Speedup Possible Curves.

VI. Results and Discussion

We evaluate the proposed parallelization technique using a dual hierarchical hybrid parallel programming model via the implementation of linear algebraic dense matrix multiplication. All results have been executed using the quad-core processor. Besides, we quantified various matrix sizes. We also examine linear DMM application performance in different datasets, assisting in determining the execution times in secs. Besides, four well-known automatic parallelization tools S2PMOACC, Cetus, Par4all, and S2P were included in the study. The speedup of studied tools was calculated to determine the efficiency of each tool.

In the 1st experiment, we run a serial matrix multiplication by different sizes. Then, we run same matrix using hybrid parallel code executing in different number of MPI processes to evaluate the execution time between serial and hybrid version. We observed the increasing in execution time in serial version unlike parallel codes. Fig. 10, Fig. 11, and Fig. 12 shows the obtained results.

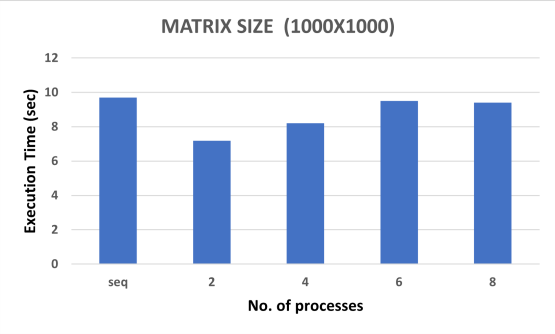


Fig. 10. Execution Time of Matrix 1000x1000.

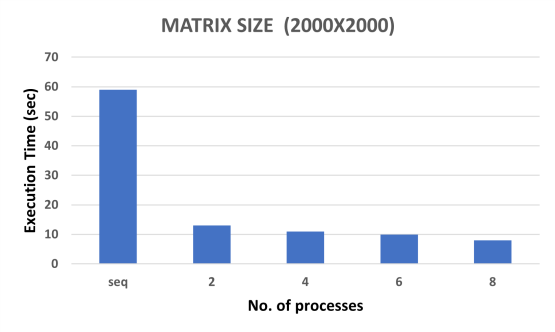


Fig. 11. Execution Time of Matrix 2000x2000.

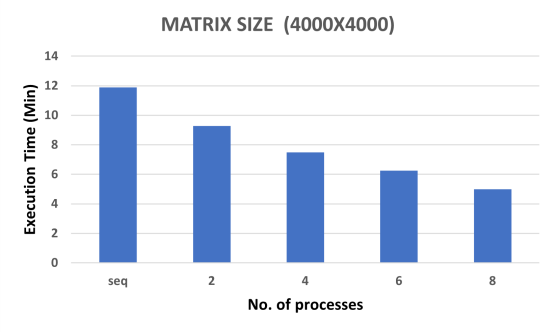


Fig. 12. Execution Time of Matrix 4000x4000.

In the 2nd experiment with a minute amount of matrix multiplication by 1000 x 1000 matrix size, we experiment this dataset with single GPU by running the existing tools along with our proposed solution. S2PMOACC take 8 secs to complete the execution while Cetus computes in 13 secs, Par4All end its execution in 22 secs, and S2P calculated executed time is 24 secs. From Fig. 13, the results show that our proposed model execution time is negligible in minimum resources. Besides, we increase the GPU cores and analyze the same translation tools with an equal number of given devices.

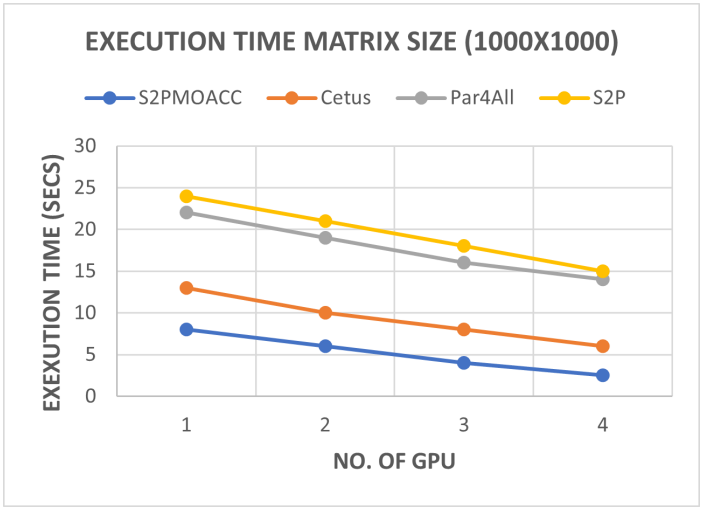


Fig. 13. Performance (Execution Time) in DMM.

According to Fig. 13, we measure second performance matric as speedup in experiment 2 that involves the optimal time taken in serial to parallel processing for all the implementation run with an equal number of resources. Fig. 14 shows the speedup of our tool along with other tools when Using 1000 x 1000 matrix multiplication with different number of GPU.

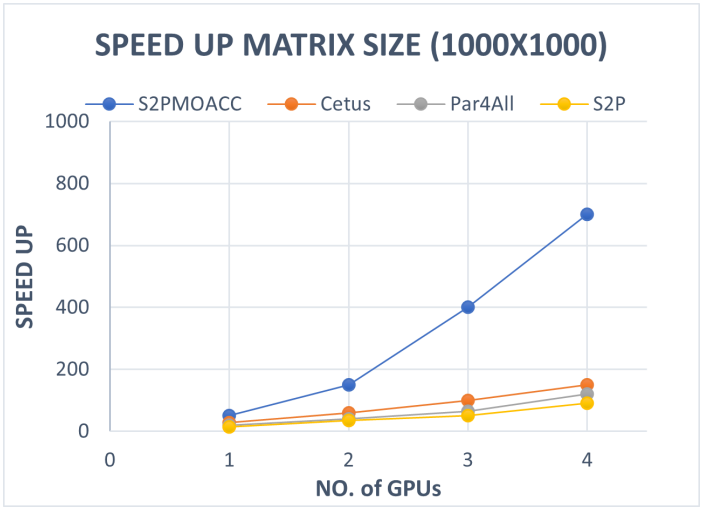


Fig. 14. Performance (Speedup) in DMM.

Finally, we study the capabilities and features of the included tools. We observed that our tool has more features than other tools. Table II provide the summary of the comparison.

TABLE II. TOOLS SPECIFICATIONS COMPARISONS.

Feature	Tools			
	S2PMOACC	**Cetus**	**Par4all**	**S2P**
Operating Systems	Windows Linux	Linux	Linux	Linux
Input Language	C,C++	C	Fortran,C	C
Technique used	MPI OpenACC	OpenMP	OpenMP CUDA	OpenMP Pthread
Hybrid Output	Yes	NO	NO	NO
Support For loop	Yes	Yes	Yes	Yes
Support While	Yes	NO	NO	NO
Support Do_while	Yes	NO	NO	NO

VII. CONCLUSION

The application of HPC has increased significantly in all scientific fields and HPC systems has been used to solve complex computational programs. Despite running software programs sequentially, researchers and programmers face difficulties in dealing with huge and complex programs, which increase the execution time. The serial code must be converted to parallel code to improve the program performance and reduce the execution time. Therefore, parallelization tools must assist programmers in the converting process. In this work, we proposed a novel automatic translation tool that converts serial C++ code into parallel code using a hybrid parallel programming model. This auto-translation tool supports a dual hierarchical MPI and OpenACC parallel computing model for heterogeneous systems that use GPU devices for providing parallelism. We implement the proposed solution in the DMM application by using different number of MPI processes in the first experiment. The second experiment compare the proposed tool execution time and speedup with well-known auto-translation tools such as Cetus, Par4all, and S2P tools. The Third experiment compare the features and limitations between tools. Based on the experimental results, the S2PMOACC outperformed the existing tools and provides complete auto parallelism in all performance metrics.

VIII. FUTURE WORK

In the near future ,the auto-parallel computing systems are in high demand as they support ECS applications. Therefore, we must have an adaptive auto-translation technique for parallelizing sequential code to support the future Exascale-computing system, large-scale cluster system, multi-core distributed system, and heterogeneous cluster system. For that, we aim to implement more enhancement to our proposed tool to keep pace with the continuous development of future systems.

REFERENCES

[1] J. Diaz, C. Muñoz-Caro, and A. Niño, "A survey of parallel programming models and tools in the multi and many-core era," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 8, pp. 1369–1386, 2012.

[2] A. Roy, J. Xu, and M. H. Chowdhury, "Multi-core processors: A new way forward and challenges," in *2008 International Conference on Microelectronics*, 2008, pp. 454–457.

[3] A. Barve, S. Khomane, B. Kulkarni, S. Ghadage, and S. Katare, "Parallelism in c++ programs targeting objects," in *2017 International Conference on Advances in Computing, Communication and Control (ICAC3)*, 2017, pp. 1–6.

[4] A. Athavale, P. Ranadive, M. Babu, P. Pawar, S. Sah, V. Vaidya, and C. Rajguru, "Automatic sequential to parallel code conversion," *GSTF Journal on Computing (JoC)*, vol. 1, no. 4, 2014.

[5] D. Akhmetova, R. Iakymchuk, O. Ekeberg, and E. Laure, "Performance study of multithreaded mpi and openmp tasking in a large scientific code," in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2017, pp. 756–765.

[6] K. R. Varsha, "Automatic parallelization tools : A review," *International Journal of Engineering Science and Computing IJESC*, vol. 7, no. 3, pp. 5780–5784, 2017.

[7] A. Barve, S. Khandelwal, N. Khan, S. Keshatiwar, and S. Botre, "Serial to parallel code converter tools: A review," in *International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue National Conference" NCPCI-2016*, vol. 19, 2016.

[8] A. M. Alghamdi, F. E. Eassa, M. A. Khamakhem, A. S. A. AL-Ghamdi, A. S. Alfakeeh, A. S. Alshahrani, and A. A. Alarood, "Parallel hybrid testing techniques for the dual-programming models-based programs," *Symmetry*, vol. 12, no. 9, p. 1555, 2020.

[9] M. Mathews and J. P. Abraham, "Automatic code parallelization with openmp task constructs," in *2016 International Conference on Information Science (ICIS)*, 2016, pp. 233–238.

[10] J. Etancelin and J. Kraus, "Multi-GPU programming with OpenACC and MPI," in *GPU Technology Conference*, 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01471165

[11] A. M. Alghamdi and F. E. Eassa, "Parallel hybrid testing tool for applications developed by using mpi+ openacc dual-programming model," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 4, no. 2, pp. 203–210, 2019.

[12] "Message passing interface forum." [Online]. Available: https://www.mpi-forum.org/

[13] P. Balaji and W. Gropp, "Advanced mpi programming," 2016. [Online]. Available: https://www.mcs.anl.gov/~thakur/sc16-mpi-tutorial/slides.pdf

[14] C.-T. Yang, C.-L. Huang, and C.-F. Lin, "Hybrid cuda, openmp, and mpi parallel programming on multicore gpu clusters," *Computer Physics Communications*, vol. 182, no. 1, pp. 266–269, 2011.

[15] "Mpich implementation." [Online]. Available: https://www.mpich.org/

[16] "Open mpi: Open source high performance computing." [Online]. Available: https://www.open-mpi.org/

[17] "Intel mpi library." [Online]. Available: https://software.intel.com/

[18] "Ibm spectrum mpi - overview — ibm." [Online]. Available: https://www.ibm.com/products/spectrum-mpi

[19] A. S. A. Alghamdi, A. M. Alghamdi, F. E. Eassa, and M. A. Khemakhem, "Acc_test: Hybrid testing techniques for mpi-based programs," *IEEE Access*, vol. 8, pp. 91 488–91 500, 2020.

[20] J. Kim, S. Lee, and J. S. Vetter, "Impacc: a tightly integrated mpi+ openacc framework exploiting shared memory parallelism," in *Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing*, 2016, pp. 189–201.

[21] R. Farber, *Parallel programming with OpenACC*. Newnes, 2016.

[22] S. Chandrasekaran and G. Juckeland, *OpenACC for Programmers: Concepts and Strategies*. Addison-Wesley Professional, 2017.

[23] M. U. Ashraf, F. A. Eassa, and A. A. Albeshri, "Efficient execution of smart city's assets through a massive parallel computational model," in *International Conference on Smart Cities, Infrastructure, Technologies and Applications*. Springer, 2017, pp. 44–51.

[24] J. Dinan, P. Balaji, D. Buntinas, D. Goodell, W. Gropp, and R. Thakur, "An implementation and evaluation of the mpi 3.0 one-sided communication interface," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 17, pp. 4385–4404, 2016.

[25] H. Jin, D. Jespersen, P. Mehrotra, R. Biswas, L. Huang, and B. Chapman, "High performance computing using mpi and openmp on multi-

core parallel systems," *Parallel Computing*, vol. 37, no. 9, pp. 562–575, 2011.

[26] D. Akhmetova, R. Iakymchuk, O. Ekeberg, and E. Laure, "Performance study of multithreaded mpi and openmp tasking in a large scientific code," in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2017, pp. 756–765.

[27] T. Katagiri, "Basics of mpi programming," in *The Art of High Performance Computing for Computational Science, Vol. 1*. Springer, 2019, pp. 27–44.

[28] D. Jacobsen, J. Thibault, and I. Senocak, "An mpi-cuda implementation for massively parallel incompressible flow computations on multi-gpu clusters," in *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, 2010, p. 522.

[29] F. Bonelli, M. Tuttafesta, G. Colonna, L. Cutrone, and G. Pascazio, "An mpi-cuda approach for hypersonic flows with detailed state-to-state air kinetics using a gpu cluster," *Computer Physics Communications*, vol. 219, pp. 178–195, 2017.

[30] S. Sah and V. G. Vaidya, "A review of parallelization tools and introduction to easypar," *International Journal of Computer Applications*, vol. 56, no. 12, 2012.

[31] S. Prema and R. Jehadeesan, "Analysis of parallelization techniques and tools," *International Journal of Information and Computation Technology*, vol. 3, no. 5, pp. 471–478, 2013.

[32] A. J. Almghawish, A. M. Abdalla, and A. B. Marzouq, "An automatic parallelizing model for sequential code using python," *International Journal*, vol. 7, no. 3, 2017.

[33] K. Alsubhi, F. Alsolami, A. Algarni, K. Jambi, F. Eassa, and M. Khemakhem, "An architecture for translating sequential code to parallel," in *Proceedings of the 2nd International Conference on Information System and Data Mining*, 2018, pp. 88–92.

[34] M. D. Hill and M. R. Marty, "Amdahl's law in the multicore era," *Computer*, vol. 41, no. 7, pp. 33–38, 2008.

[35] S. Ristov, R. Prodan, M. Gusev, and K. Skala, "Superlinear speedup in hpc systems: Why and when?" in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2016, pp. 889–898.

[36] C. A. Navarro, N. Hitschfeld-Kahler, and L. Mateu, "A survey on parallel computing and its applications in data-parallel problems using gpu architectures," *Communications in Computational Physics*, vol. 15, no. 2, pp. 285–329, 2014.

Fully Convolutional Networks for Local Earthquake Detection

Youness Choubik[1]
LIMIARF Laboratory
Faculty of Sciences
Mohammed V University
in Rabat, Morocco

Abdelhak Mahmoudi[2]
Ecole Normale Supérieure
Mohammed V University
in Rabat, Morocco

Mohammed Majid Himmi[3]
LIMIARF Laboratory
Faculty of Sciences
Mohammed V University
in Rabat, Morocco

***Abstract*—Automatic earthquake detection is widely studied to replace manual detection, however, most of the existing methods are sensitive to seismic noise. Hence, the need for Machine and Deep Learning has become more and more significant. Regardless of successful applications of the Fully Convolutional Networks (FCN) in many different fields, to the best of our knowledge, they are not yet applied in earthquake detection. In this paper, we propose an automatic earthquake detection model based on FCN classifier. We used a balanced subset of STanford EArthquake Dataset (STEAD) to train and validate our classifier. Each sample from the subset is re-sampled from 100Hz to 50Hz then normalized. We investigated different, widely used, feature normalization methods, which consist of normalizing all features in the same range, and we showed that feature normalization is not suitable for our data. On the contrary, sample normalization, which consists of normalizing each sample of our dataset individually, improved the accuracy of our classifier by ∼16% compared to using raw data. Our classifier exceeded 99% on training data, compared to ∼83% when using raw data. To test the efficiency of our classifier, we applied it to real continuous seismic data from XB Network from Morocco and compared the results to our catalog containing 77 earthquakes. Our results show that we could detect 75 out of 77 earthquakes contained in the catalog.**

***Keywords*—Earthquake detection; fully convolutional networks; data normalization; classification**

I. Introduction

Earthquake detection requires discriminating real earthquakes from noise signals, which makes it a classification problem. Earthquake detection is a very crucial and challenging phase in seismic processing, especially for single station-based detection, because every station records a very wide range of non-earthquake waveforms. Manual detection is a time consuming work due to the huge amount of seismic data, therefore, automatic earthquake detection is essential and widely studied.

A large number of automatic earthquake detection methods exist [35], some of them are time domain methods, such as the short term average to long term average (STA/LTA), which is the most used in seismic stations. Other time domain methods are used, such as the maximum likelihood detector [9], envelope-based detector [3], and modified data envelope detector [29]. On the other hand, some frequency domain methods are based on the Power Spectral Density (PSD) [25] and the Walsh transform [12]. However, most of the existing methods are sensitive to noise and suffer from false and missed detections [32].

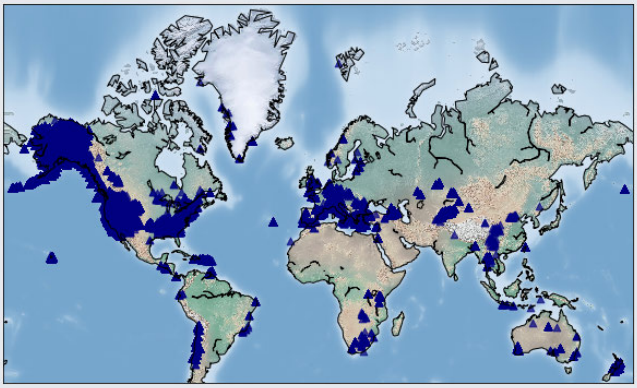
In recent years, methods based on Machine and Deep Learning have shown great potentials, especially Artificial Neural Networks, which are widely used in seismic detection [8], [10], [1]. The Convolutional Neural Networks (CNNs) known as very successful in the computer vision area become more and more popular in seismic area [23], [39], [37], [34]. Recurrent Neural Networks (RNNs), another architecture of Neural Networks known to be suitable for many time-series applications such as text to speech and voice recognition [36], are also used in seismology [40], [19], [6].

Unsupervised clustering methods are also used in seismology. They can cluster seismic samples into different clusters without prior knowledge of labels. Different clustering methods are used in many seismic studies, such as k-means [5], [28], Deep Convolutional Autoencoders [21], and Self-Organizing Maps [16], [17], [27].

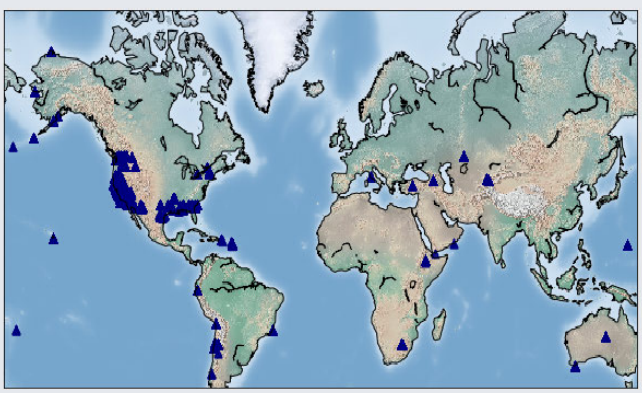
Fully Convolutional Networks (FCN) [24] are a Neural Network architectures that have been successfully applied in many different fields, such as image segmentation[13], [4], medical image analysis [7], [18], character recognition [31], time-series classification [33], [14], [22] and also in seismology; for earthquake localization, by taking a window of three-component waveform data from multiple stations and predicting the earthquake location with a 3D image [38], and for fault detection, where the FCN model extracts fault features from synthetic seismic data and recognize the locations of faults with an accuracy of ∼97% [26]. Despite their higher achievements, to the best of our knowledge, FCN had not yet been applied to seismic detection.

In this study, we describe the application of FCN for earthquake detection using seismic waveforms from a single seismic station. The basic of the earthquake detection problem is turned into a classification problem by using a subset of STanford EArthquake Dataset (STEAD) to train our classifier. Our approach does not require a feature extraction technique, which makes it independent of the choice of sensitive features. We tested the effectiveness of our classifier by applying it to real continuous data From the XB seismic network implemented in morocco between 2009 and 2013 [2].

In the following, we first describe the dataset used to train our classifier and the real continuous dataset used for testing. Then, we present the method and the steps applied in the



(a) Locations of seismic stations recording earthquakes



(b) Locations of seismic stations recording seismic noise

Fig. 1. Distribution of Seismic Stations used to Record Earthquakes (a) and Noise (b) [20].

training process. Finally, we describe the results and discuss the performance of our classifier Using real continuous seismic data.

II. Data Description

The STanford EArthquake Dataset (STEAD) [20] is a large-scale and global dataset that contains two waveform classes; seismic noise and local earthquake waveforms, which are recorded at local distances (within 350 km of earthquakes). STEAD comprises about 1.2 million waveforms, recorded by worldwide located seismometers, resampled at 100Hz, and have 60 seconds duration (6000 features). Local-earthquakes class contains about 1 050 000 three-component seismograms associated with ∼450 000 earthquakes that occurred between January 1984 and August 2018 (Fig. 1(a)). The seismic noise class contains about 100 000 waveforms that have been recorded since 2000 in the United States and Europe (Fig. 1(b)).

The earthquake waveforms are requested from continuous time-series archived at the Incorporated Research Institutions for Seismology Data Management Center (IRIS DMC). Three types of arrival statuses exist, “Manual” picks; picked manually by human analysts, “automatic” picks; measured by automatic algorithms and “autopicker” that are determined using an AI-based model. STEAD is provided as individual arrays containing three waveforms that correspond to three-component seismograms, each waveform has 6000 features.

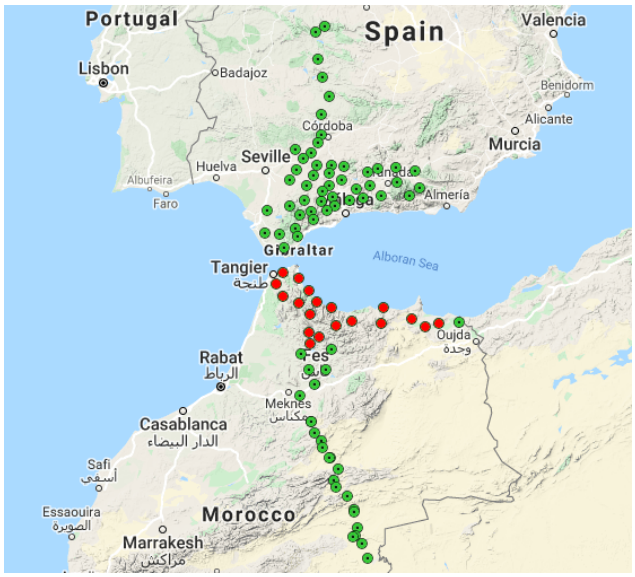


Fig. 2. XB Seismic Network Installed in both Morocco and Spain, it Contains 93 Seismic Stations, from which 19 Stations are used in this Study and are Colored in Red.

The XB network used to test our model was deployed in both Morocco and Spain, in the frame of the Project to Investigate Convective Alboran Sea System Overturn (PICASSO), from 2009 to 2013. The XB network contained 93 seismic stations labeled as PICASSO Morocco (PM) and PICASSO Spain (PS). Fig. 2 shows the stations of the XB network, where 44 stations were installed in Morocco. We used data of January 2011 from 19 stations, measured by High gain Broadband (BH) seismic instruments and sampled at 50Hz, to test our model.

III. Training with the Fully Convolutional Networks

To train our model, we choose a subset of STEAD measured by BH seismic instruments, since we have only BH waveforms from XB network. We found 7874 unique noise waveforms of BH type in STEAD. In order to create a balanced dataset, we extracted the same quantity of waveforms from the earthquake class. because classification is affected by imbalanced datasets and resulting a reduction in accuracy as shown by [30]. The selected waveforms are associated with a wide range of earthquake sizes from magnitude 0 to magnitude 6.3. Earthquakes were recorded within 330 km of the earthquakes, are mainly shallower than 210 km and have Signal to Noise Ratio (SNR) between -5 and 100 decibels.

Our dataset is comprised of 15 748 samples and divided into train/validation/test subsets as shown in Table I. The portion of the test-set is small because we will test our model on real continuous data from the XB seismic network. The

TABLE I. TRAIN/VALIDATION/TEST SUBSETS DISTRIBUTION USED IN OUR STUDY

Training-set	Validation-set	Test-set
12000 (6000 from earthquake class, 6000 from noise class)	3000 (1500 from earthquake class, 1500 from noise class)	748 (374 from earthquake class, 374 from noise class)

samples used in STEAD are 60 seconds waveforms sampled at 100Hz. Since we are applying our model to data from XB network that is sampled at 50Hz, we resampled our dataset to 50Hz, so that every sample have 3000 features instead of 6000.

The Fully Convolutional Network classifier used in this study is comprised of four convolutional layers with different filter numbers and sizes (Fig. 3), followed by batch normalization that normalizes the output of the convolution layer and a ReLU activation function, which enables better training of deeper networks, compared to other activation functions [11], then a Global Pooling layer that reduces the amount of parameters in the network to an output prediction for the model. Finally, since the output is One Hot Encoded, a softmax function is placed in the output layer that normalizes the output into two probabilities corresponding to belonging to the two classes earthquake and noise. The adaptive moment estimation algorithm (Adam) is used as optimizer for our classifier.

The classifier is trained to distinguish between earthquake and noise signals using the STEAD subset described above. The training/validation subsets were randomly split using a 5-fold cross-validation. The training was performed on 100 epochs, where each epoch is a complete pass through the entire training dataset, with early stopping enabled, which stop the training when the loss (error) does not decrease during training. We used a learning rate decay, where the learning rate is reduced by a factor of 10 once learning stagnates for a number of epochs. The predictions were compared to the real classes then the loss and accuracy are calculated.

Normalization is one of the most used data preparation techniques in deep learning, because features often have different ranges of values, which make the training process takes a long time to converge. Feature normalization and standardization are the most used methods. To select the best normalization method, we compared different methods against raw data (without applying any normalization method) and selected the one that gave the best accuracy on the training/validation data. The methods applied to our input data are the following:

- MinMax: Transform features by scaling each feature individually between zero and one.
- MaxAbs: Scale each feature by its maximum absolute value such that the maximal absolute value of each feature in the training set will be 1.0.
- Standard: Standardize features by removing the mean and scaling to unit variance.
- RobustScaler: Removes the median and scales the data according to the quantile range independently on each feature.

IV. RESULTS AND DISCUSSIONS

In this section, we will present and discuss the effect of using different normalization methods and batch sizes on the classifier accuracy. We investigated the effect of normalizing features on the model accuracy and compared it against using raw data. We conducted many experiments using the same training process for the normalization methods described above. Table II shows the mean accuracy of 5-fold cross-validation. We can see that MinMax, MaxAbs and Robust normalizations decrease the model accuracy compared to raw data, while standardization improves slightly the accuracy. Overall, we see that normalizing the input features did not bring a big improvement to our model, so we suspected the feature normalization to be not suitable for our data. Therefore, we tried to normalize our data per sample instead of normalizing per feature. By normalizing per sample, we mean that each sample of our dataset is normalized individually. We reported the results in Table III.

By using sample-normalized data, we can clearly see an improvement of the accuracy compared to feature-normalized and raw data. All the methods improved the accuracy without exception, compared to feature-normalized, especially the MinMax method, which is improved by $\sim$35%. The sample-standardization method made the best accuracy over the other methods, it reaches 99% on the training data, with an improvement of $\sim$16% and $\sim$14% compared to raw data feature-standardization respectively.

As seen in Fig. 4(b), when standardizing per feature, the range of earthquake classes is very large compared to that of noise classes, which makes no difference with raw data (Fig. 4(a)), except for the scale of the signals. It can be observed from Fig. 4(c) that both earthquake and noise samples have close ranges when standardized per sample. Hence, the classifier is forced to classify samples based on their shape instead of their amplitude. In the rest of our tests, only the sample-standardization method will be presented, since it outperformed the other methods.

Different batch sizes are investigated, where each batch is a subset of signals given to the network at once. Fig. 5 shows an example of the evolution of the loss function during the training process and it is clear that our classifiers converge as the training progresses. We can see that for larger batch sizes, the training loss is bigger and the validation loss is smoother, because large batch sizes are less sensitive to outliers, and converge slower than small batch sizes as stated by other studies [15].

Fig. 6(a) shows the accuracy during the training process, we can clearly see that larger batch sizes have lower accuracies compared to smaller batch sizes. While for the validation dataset (Fig. 6(b)), larger batch sizes tend to be slower and

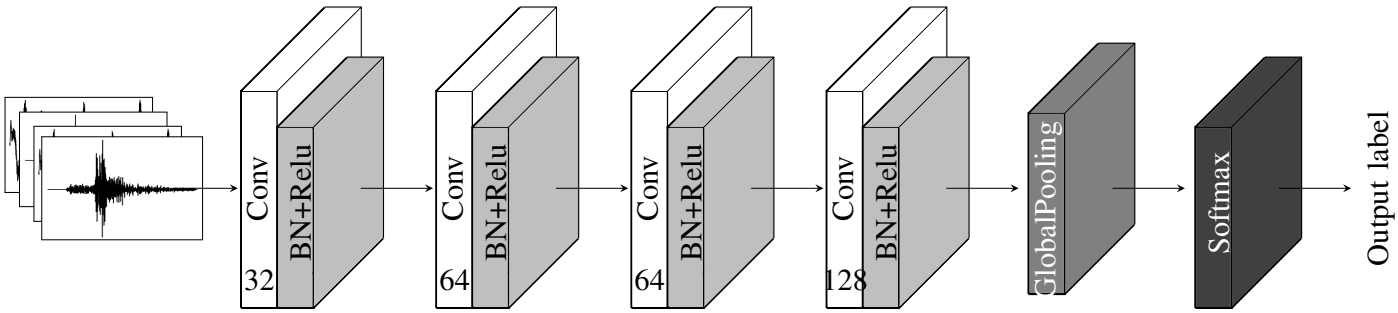


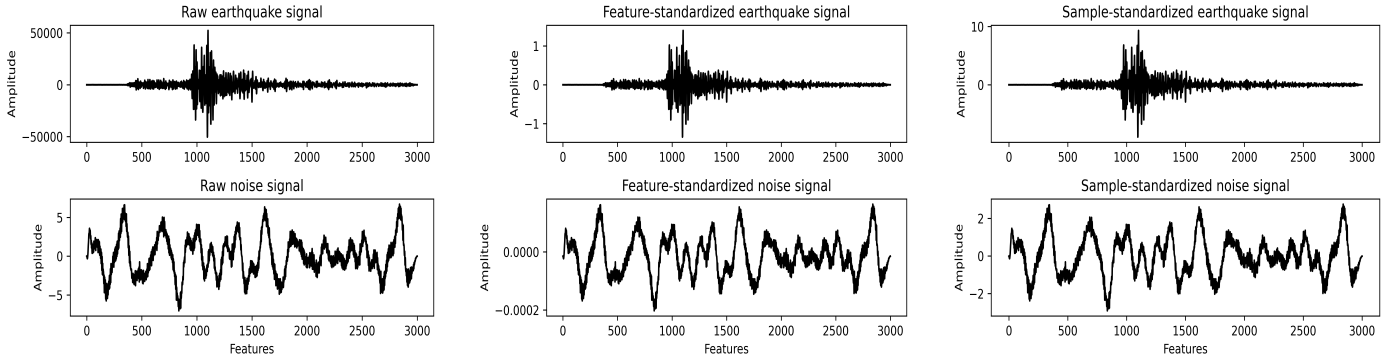
Fig. 3. The FCN Architecture Consists of 4 Convolutional Layers Followed by Batch-Normalization (BN) then a Rectified Linear Units (ReLU) Activation Function, Finally Global-Pooling and Softmax Layers.

TABLE II. THE ACCURACY FOR DIFFERENT FEATURE NORMALIZATION METHODS

	Raw data	Standard	MinMax	MaxAbs	Robust
Training	0.838	0.858	0.639	0.822	0.778
Validation	0.785	0.885	0.619	0.683	0.718
Test	0.611	0.894	0.502	0.513	0.657

TABLE III. THE ACCURACIES FOR DIFFERENT PER-SAMPLE NORMALIZATION METHODS

	Raw data	Standard	MinMax	MaxAbs	Robust
Training	0.838	0.990	0.986	0.985	0.982
Validation	0.785	0.987	0.985	0.986	0.916
Test	0.611	0.998	0.995	0.993	0.973

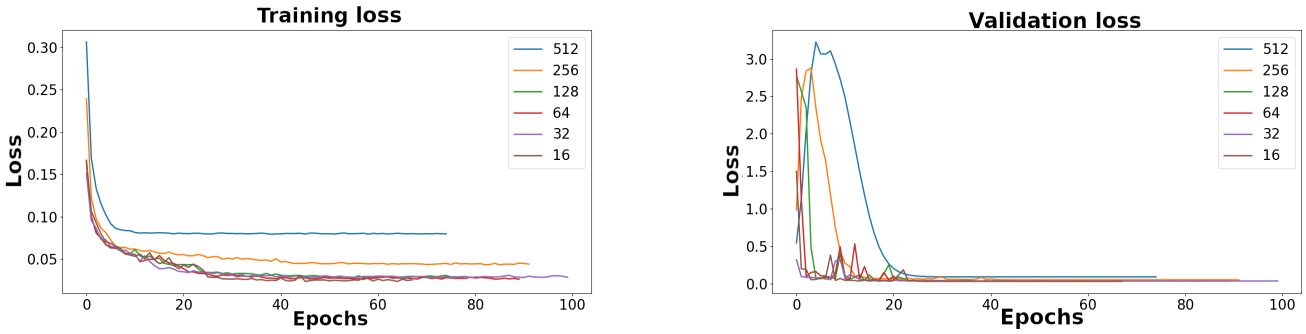


(a) Raw signal

(b) Normalize per feature

(c) Normalize per sample

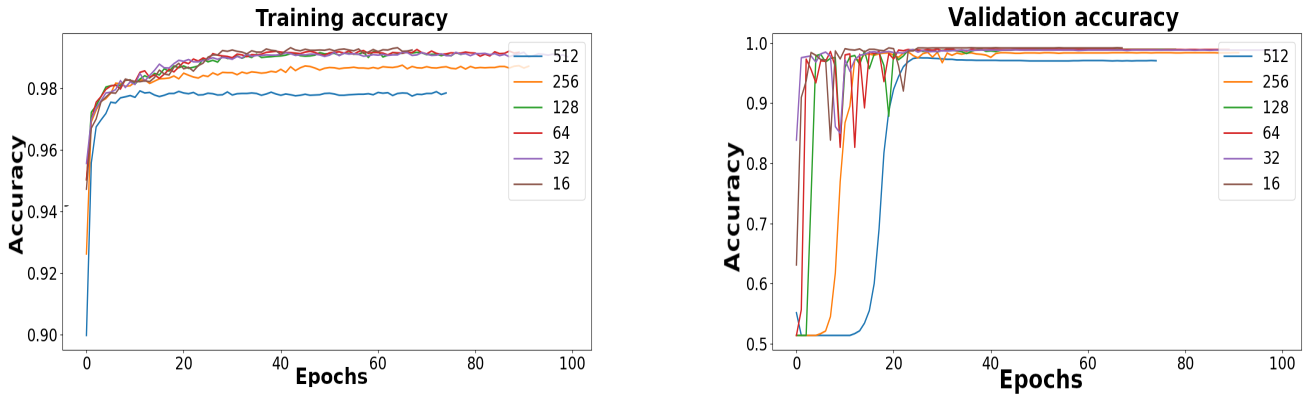
Fig. 4. Standardization Effect for the Same Two Signals, in Top an Earthquake Signal while in Bottom a Noise One. Figure (a) Shows Raw Signals, in (b) Both Signals are Standardized Per Feature, (c) Shows the Sample-Standardization Version for Both Signals.



(a) Training loss

(b) Validation loss

Fig. 5. Model Loss Per Batch Size.



(a) Training accuracy

(b) Validation accuracy

Fig. 6. Training and Validation Subsets Accuracies Per Batch Size. Large Batches Tend to have Lower Accuracies for Both Subsets.

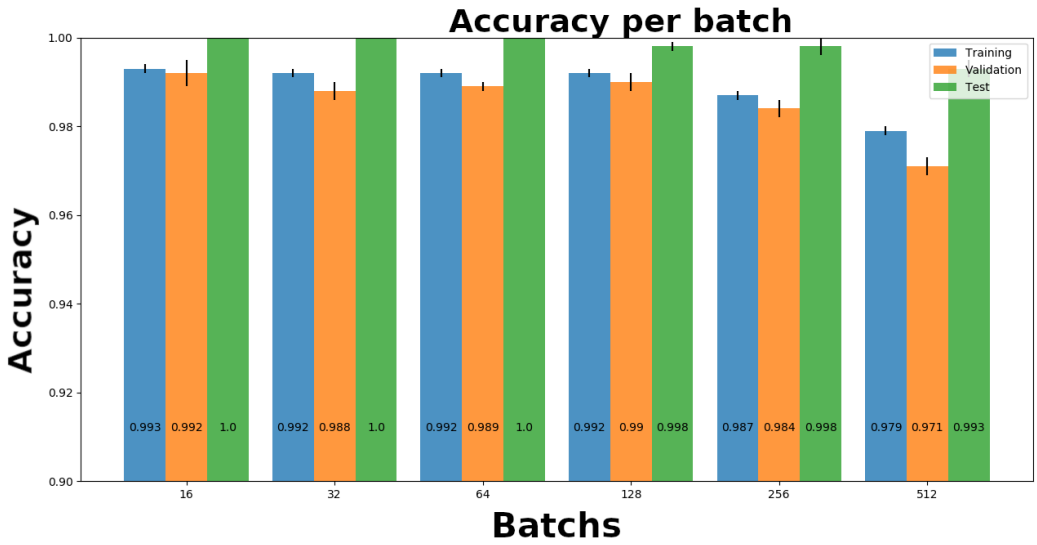


Fig. 7. The Mean Accuracy Per Batch Size. The Best Accuracies in Train/Validation/Test Subsets are Obtained by using a 16 Batch Size.

more stable. Fig. 7 shows the mean accuracy of 5-fold cross-validation. The best accuracies in training, validation and test are 99.3%, 99.2% and 100% respectively, obtained by using a 16 batch size. For smaller batches, the accuracy in training reached 100%, but in validation it has fallen to 70%, which means that the model over-fit and can not generalize for new data. The high accuracy in test-set is due to the small amount of data, because we are interested in testing our classifier on continuous data from XB network.

To check the effectiveness of our best classifier, we tested it on real three-component seismic data from the XB network. The test was applied to data from the first month of 2011, from 19 seismic stations, presented in red in fig. 2. The frequency of the seismic data is about 50Hz, and the input feature, which will be fed to the classifier, is a sliding window of 60 seconds length (3000 features), and the window is moved by 15 sec after each test.

To verify our results, we compared the earthquakes detected by our classifier to a seismic catalog that we have. Our catalog contains 77 earthquakes of magnitude > 2, located in the region of XB network. By comparing our results with the catalog, we found that our classifier detected 75 out of the 77 earthquakes contained in the catalog. Our analysis shows that our classifier is able to reliably detect local earthquake signals in continuous real data.

V. Conclusion

In this paper, we have presented a seismic detection model, based on a Fully Convolutional Networks classifier which is trained on STanford EArthquake Dataset (STEAD) and tested on real continuous seismic data. By making a separate standardization for each sample of our dataset, instead of normalizing per feature, the performance of our classifier is increased significantly by $\sim$16% compared to raw data. Our experiments show that the use of small batch sizes is more adequate for our dataset, however, very small batch sizes (8 and lower) make the model over-fit and can not generalize for new data. By applying our classifier to real continuous data from XB network in Morocco, we were able to detect local earthquakes already existing in our catalog. Our method does not require hand-engineered features and is able to discriminate between earthquakes and seismic noise with high accuracy. Our results demonstrated that FCN classifier holds vast promise for making seismic detection more accurate.

References

[1] Jubran Akram, Oleg Ovcharenko, and Daniel Peter. A robust neural network-based approach for microseismic event detection. In *SEG Technical Program Expanded Abstracts 2017*, pages 2929–2933, 08 2017.

[2] Levander Alan and Humphreys Gene. Program to investigate convective alboran sea system overturn, 2009.

[3] Rex Allen. Automatic earthquake recognition and timing from single trace. *Bulletin of the Seismological Society of America*, 68:1521–1532, 10 1978.

[4] Lei Bi, Jinman Kim, Euijoon Ahn, Ashnil Kumar, David Dagan Feng Feng, and Michael Fulham. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. *Pattern Recognition*, 85:78–89, 01 2019.

[5] Yangkang Chen, Guoyin Zhang, Min Bai, Shaohuan Zu, Zhe Guan, and Mi Zhang. Automatic waveform classification and arrival picking based on convolutional neural network. *Earth and Space Science*, 6, 04 2019.

[6] Jana Doubravová, Jan Wiszniowski, and Josef Horalek. Single layer recurrent neural network for detection of swarm-like earthquakes in w-bohemia/vogtland - the method. *Computers & Geosciences*, 93, 05 2016.

[7] Jingfan Fan, Xiaohuan Cao, and Pew-Thian Yap. Birnet: Brain image registration using dual-supervised fully convolutional networks. *Medical Image Analysis*, 54, 02 2018.

[8] Luigi Fortuna, Salvatore Graziani, M. Presti, and Giuseppe Nunnari. A neural network for seismic events classification. In *[Proceedings] IGARSS'91 Remote Sensing: Global Monitoring for Earth Management*, pages 1663 – 1666, 07 1991.

[9] Walter Freiberger. An approximate method in signal detection. *Quarterly of Applied Mathematics*, 20, 01 1963.

[10] Flora Giudicepietro, Anna Esposito, and Patrizia Ricciolino. Fast discrimination of local earthquakes using a neural approach. *Seismological Research Letters*, 88:1089–1096, 07 2017.

[11] Xavier Glorot, Antoine Bordes, and Y. Bengio. Deep sparse rectifier neural networks. *Proceedings of the 14th International Conference on Artificial Intelligence and Statisitics (AISTATS) 2011*, 15:315–323, 01 2011.

[12] Tom Goforth. An automatic signal detection algorithm based on the walsh transform. *Bulletin of the Seismological Society of America*, 71:1351, 01 1981.

[13] Yuan Huang, Fugen Zhou, and Jerome Gilles. Empirical curvelet based fully convolutional network for supervised texture image segmentation. *Neurocomputing*, 349, 04 2019.

[14] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Lstm fully convolutional networks for time series classification. *IEEE Access*, PP, 09 2017.

[15] Nitish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tang. On large-batch training for deep learning: Generalization gap and sharp minima. 09 2016.

[16] Andreas Köhler, Matthias Ohrnberger, and Frank Scherbaum. Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps. *GEOPHYSICAL JOURNAL INTERNATIONAL*, 182:1619–1630, 09 2010.

[17] Andreas Köhler, Matthias Ohrnberger, Carsten Riggelsen, and Frank Scherbaum. Unsupervised feature selection for pattern search in seismic time series. In *Proceedings of the 2008 International Conference on New Challenges for Feature Selection in Data Mining and Knowledge Discovery - Volume 4*, FSDM'08, page 106–120. JMLR.org, 2008.

[18] Chao Li, Xinggang Wang, Wenyu Liu, Longin Jan Latecki, Bo Wang, and Junzhou Huang. Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Medical Image Analysis*, 53, 02 2019.

[19] Men-Andrin Meier, Zachary Ross, Anshul Ramachandran, Ashwin Balakrishna, Suraj Nair, Peter Kundzicz, Zefeng Li, Jennifer Andrews, Egill Hauksson, and Yisong Yue. Reliable real-time seismic signal/noise discrimination with machine learning. *Journal of Geophysical Research: Solid Earth*, 12 2018.

[20] S.Mostafa Mousavi, Yixiao Sheng, Zhu Weiqiang, and Gregory Beroza. Stanford earthquake dataset (stead): A global data set of seismic signals for ai. *IEEE Access*, PP:1–1, 10 2019.

[21] S.Mostafa Mousavi, Zhu Weiqiang, William Ellsworth, and Gregory Beroza. Unsupervised clustering of seismic signals using deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, PP:1–5, 05 2019.

[22] Kaushal Paneri, Vishnu Tv, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. Regularizing fully convolutional networks for time series classification by decorrelating filters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:10003–10004, 07 2019.

[23] Thibaut Perol, Michaël Gharbi, and Marine Denolle. Convolutional neural network for earthquake detection and location. *Science Advances*, 4, 02 2017.

[24] Evan Shelhamer, Jonathon Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1–1, 05 2016.

[25] Shensa. The deflection detector: its theory and evaluation on short-period seismic data. technical report. 01 1977.

[26] Y. Shi, L. Huang, X. Dong, T. Liu, and J. Ning. Application of Fully Convolutional Neural Network on Fault Detection. In *AGU Fall Meeting Abstracts*, volume 2018, pages S13B–04, December 2018.

[27] Chengyun Song, Zhining Liu, Yaojun Wang, Xingming Li, and Guangmin Hu. Multi-waveform classification for seismic facies analysis. *Computers & Geosciences*, 101:1–9, 04 2017.

[28] Chengyun Song, Zhining Liu, Yaojun Wang, Feng Xu, Xingming Li, and Guangmin Hu. Adaptive phase k-means algorithm for waveform classification. *Exploration Geophysics*, 49, 01 2017.

[29] S.W. Stewart. Real time detection and location of local seismic events in central california. *Bulletin of the Seismological Society of America*, 67(2):433–452, 04 1977.

[30] Mustafa Üstüner, Fusun Balik Sanli, and Saygin Abdikan. Balanced vs imbalanced training data: Classifying rapideye data with support vector machines. volume XLI-B7, 07 2016.

[31] Felipe Such, Suhas Pillai, Frank Brockler, Vatsala Singh, Paul Hutkowski, and Raymond Ptucha. Intelligent character recognition using fully convolutional neural networks. *Pattern Recognition*, 88, 12 2018.

[32] Yoones Vaezi and Mirko Baan. Comparison of the sta/lta and power spectral density methods for microseismic event detection. *Geophysical Journal International*, 203:1896–1908, 12 2015.

[33] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1578–1585, 05 2017.

[34] Andy Wilkins, Andrew Strange, Yi Duan, and Xun Luo. Identifying microseismic events in a mining scenario using a convolutional neural network. *Computers & Geosciences*, 137:104418, 04 2020.

[35] Mitchell Withers, Richard Aster, Christopher Young, Judy Beiriger, Mark Harris, Susan Moore, and Julian Trujillo. A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bulletin of the Seismological Society of America*, 88:95–106, 02 1998.

[36] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31:1–36, 05 2019.

[37] Sanyi Yuan, Jiwei Liu, Shangxu Wang, Tieyi Wang, and Peidong Shi. Seismic waveform classification and first-break picking using convolution neural networks. *IEEE Geoscience and Remote Sensing Letters*, PP:1–5, 01 2018.

[38] Xiong Zhang, Jie Zhang, Congcong Yuan, Sen Liu, Zhibo Chen, and Weiping Li. Locating induced earthquakes with a network of seismic stations in oklahoma via a deep learning method. *Scientific Reports*, 10, 12 2020.

[39] Bendong Zhao, Shanzhu Xiao, Huanzhang Lu, and Junliang Liu. Waveforms classification based on convolutional neural networks. In *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 162–165, 03 2017.

[40] Jing Zheng, Jiren Lu, Suping Peng, and Tianqi Jiang. An automatic microseismic or acoustic emission arrival identification scheme with deep recurrent neural networks. *Geophysical Journal International*, 212:1389–1397, 02 2018.

A Hybridized Deep Learning Method for Bengali Image Captioning

Mayeesha Humaira[1], Shimul Paul[2], Md Abidur Rahman Khan Jim[3], Amit Saha Ami[4], Faisal Muhammad Shah[5]
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

***Abstract*—An omnipresent challenging research topic in computer vision is the generation of captions from an input image. Previously, numerous experiments have been conducted on image captioning in English but the generation of the caption from the image in Bengali is still sparse and in need of more refining. Only a few papers till now have worked on image captioning in Bengali. Hence, we proffer a standard strategy for Bengali image caption generation on two different sizes of the Flickr8k dataset and BanglaLekha dataset which is the only publicly available Bengali dataset for image captioning. Afterward, the Bengali captions of our model were compared with Bengali captions generated by other researchers using different architectures. Additionally, we employed a hybrid approach based on InceptionResnetV2 or Xception as Convolution Neural Network and Bidirectional Long Short-Term Memory or Bidirectional Gated Recurrent Unit on two Bengali datasets. Furthermore, a different combination of word embedding was also adapted. Lastly, the performance was evaluated using Bilingual Evaluation Understudy and proved that the proposed model indeed performed better for the Bengali dataset consisting of 4000 images and the BanglaLekha dataset.**

***Keywords*—Bengali image captioning; hybrid architecture; InceptionResNet; Xception**

I. Introduction

An image is worth a thousand stories. It is effortless for humans to describe these stories but it is troublesome for a machine to portray them. To obtain captions from images it is necessary to combine computer vision and natural language processing. Previously lots of research has been done on image captioning but most of them were done in English. Research done on Image captioning using other languages [13], [15], [16] is still limited. Few works until now have been conducted on image captioning in Bengali [5], [23], [37] so we aim to explore image captioning in the Bengali language further.

About 215 million people worldwide speak in Bengali among those 196 million individuals are natives from India and Bangladesh. Bengali is the 7th most utilized language worldwide[1].As a result, it is momentous to generate image captions in Bengali alongside English. Moreover, most of the natives have no knowledge of English. Additionally, image captioning can be used to aid blind people by converting the text into speech blind people who can understand the image. Also, surveillance footage can be captioned in real-time so that theft, crime or accidents can be detected faster.

The main issue of image captioning in the Bengali language is the availability of a dataset. Most of the datasets available are in English. English datasets can be translated using manual labor or using machine translation. At any rate, manual translations have higher accuracy, they are extremely monotonous and troublesome. Machine translation on the other hand provides a better solution. In our experiment, we used a Machine translator such as Google translator[2] to translate English captions to Bengali and modified those sentences that were syntactically incorrect manually. Furthermore, we also utilized BanglaLekha[3] dataset which is the only publicly available Bengali dataset for image captioning till now. All the captions in this dataset were in Bengali and human annotated. We employed two approaches to captioning images in Bengali. Firstly, a hybrid model was used as demonstrated in Fig. 1 where two embedding layers were concatenated. Among those concatenated embedding one was GloVe [22] which utilize a pre-trained file in Bengali and another was fastText [7] which was trained on the vocabulary available. Secondly, two different models were trained to have a single embedding. One was conducted with only a trainable fastText embedding and the other experimented on GloVe embedding which was pre-trained in Bengali. For all three of the cases, InceptionResnetV2 [28] and Xception [38] was used as a Convolution Neural Network (CNN) to detect objects from images.

In this work, we proposed a hybridized Deep Learning method for Image captioning. This was achieved by concatenating two word embedding. The contribution of this paper is as follows:

- We introduced a hybridized method of image captioning where two word embedding pre-trained GloVe and fastText were concatenated.
- Experiments were carried on both our models using Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU). BiGRU has not been used before for image captioning using different languages other than English.
- Moreover, these two models have been tested on two Flickr8k datasets of varying lengths. One dataset contains 4000 images and the other contains 8000 images. To our best knowledge, no paper used Flickr8k full dataset translated in Bengali for image captioning.
- Additionally, our model was also tested on the BanglaLekha dataset which contains 9154 images.

[1]https://www.vistawide.com/languages/top_30_languages.htm

[2]https://translate.google.com/

[3]https://data.mendeley.com/datasets/rxxch9vw59/2

- Lastly, it was shown that our proposed hybrid model achieved higher BLEU scores for both the Flickr4k-BN dataset and the BanglaLekha dataset.

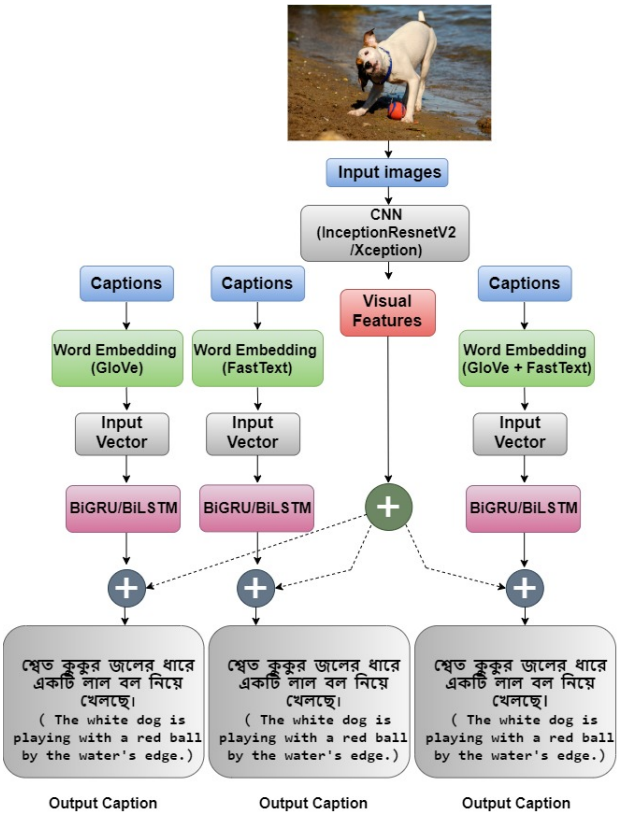


Fig. 1. Illustration of Hybridized (Right) Model and Model with Single Embedding FastText or GloVe (Left).

II. Related Work

This section depicts the progress in image captioning. Hitherto, many kinds of research have been conducted and many models have been developed to get captions that are syntactically corrected. The authors in [2] presented a model that deems the probabilistic distribution of the next word using previous word and image features. On the other hand, H. Dong et al. [6] proposed a new training method Image-Text-Image which amalgamate text-to-image and image-to-text synthesis to revamp the performance of text-to-image synthesis. Furthermore, J. Aneja [21] and S. J. Rennie [25] adapted the attention mechanism to generate caption. For vision part of image captioning Vgg16 were used by most of the papers [2], [11], [24], [25], [27], [30] as CNN but some of them also used YOLO [9], Inception V3 [6], [31], AlexNet [24], [30] ResNet [11], [18], [24] or Unet [4] as CNN for feature extraction. Concurrently, LSTM [6], [9], [11], [17], [31] was used by most of the papers for generating the next word in the sequence. However, some of the researcher also utilized RNN [19] or BiLSTM [4], [30]. Moreover, P. Blandfort et al. [32] systematically characterize diverse image captions that appear "in the wild" in order to understand how people caption images naturally. Alongside English researchers also generated captions in Chinese [15], [16], Japanese [1], Arabic [12], Bahasa Indonesia [13], Hindi [26] German [29] and Bengali [5], [23]. M. Rahman et al. [23] generated image caption in Bengali for the first time followed by T. Deb et al. [5]. Researchers of paper [23] used VGG-16 to extract image features and stacked LSTMs. On the contrary, researchers of paper [5] generated image caption using InceptionResnetV2 or VGG-16 and LSTM. They utilized 4000 images of the Flickr8k dataset to generate captions. We modified the merge model adapted by paper [5] to get much better and fluent captions in Bengali.

Only three works have been done on image captioning in Bengali till now. In [23], author's first paper, was where in image captioning in Bengali followed by [5] and [37]. Rahman et al. [23] have aimed to outline an automatic image captioning system in Bengali called 'Chittron'. Their model was trained to predict Bengali caption from input image one word at a time. The training process was carried out on 15700 images of their own dataset BanglaLekha. In their model Image feature vector and words converted to vectors after passing them through the embedding, the layer was fed to the stacked LSTM layer. One drawback of their work was that they utilized sentence BLEU score instead of Corpus BLEU score. On the other hand, Deb et al. [5] illustrated two models Par-Inject Architecture and Merge Architecture for image captioning in Bengali. In the Par-Inject model image, feature vectors were fed into intermediate LSTM and the output of that LSTM and word vectors were combined and fed to another LSTM to generate caption in Bengali. Whereas, in the Merge model image feature vectors and words vector were combined and passed to an LSTM without the use of an intermediate LSTM. They utilized 4000 images of the Fickr8k dataset and the Bengali caption their models generated were not fluent. Paper [37] used a CNN-RNN based model where VGG-16 was used as CNN and LSTM with 256 channels was used as RNN. They trained their model on the BanglaLekha dataset having 9154 images.

To overcome the above mentioned drawbacks of fluent captions we conducted our experiment using a hybridized approach. Moreover, we used 8000 images of the Flickr8k dataset alongside the Flickr4k dataset. We further validated the performance of our model using the human annotated BanglaLekha dataset.

III. Our Approach

We employed an Encoder-Decoder approach where both InceptionResnetV2 and Xception were used separately in different experimental setups to Encode Images to feature vectors and different word embedding were used to convert vocabulary to word vectors. Image feature vectors and word vectors after passing through a special kind of RNN were merged and passed to a decoder to predict captions word by word this process is illustrated in Fig. 2. We propose a hybrid model that consists of two embedding layers unlike the merge model [5]. We also conducted experiments on the merged model having either pre-trained GloVe [22] or trainable fastText [7] embedding. To be more precise, we trained the merge model using three settings as shown in Fig. 1.

Our proposed hybrid model is shown in Fig. 3. It consists of two part which is encoder and decoder.

- Encoder
  The encoder comprised of two parts one for han-

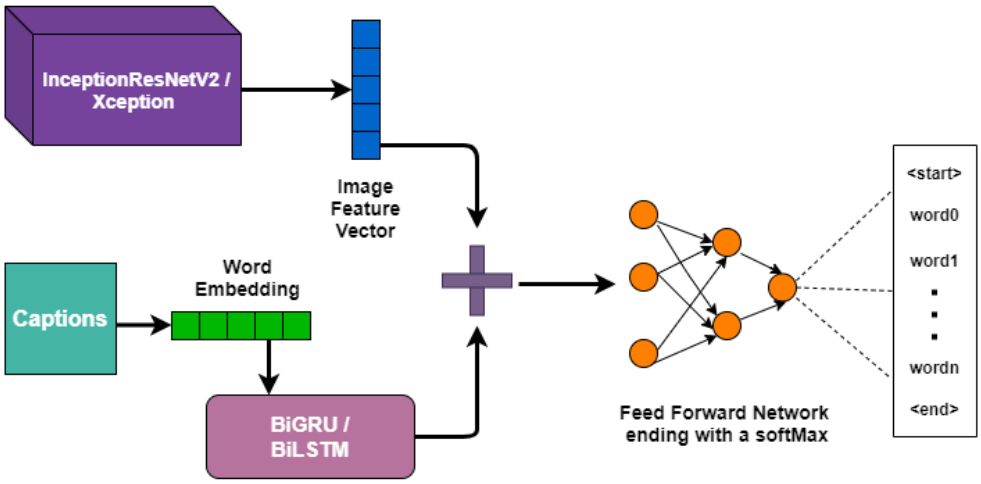


Fig. 2. An overview of how captions are generated word by word using our model.

dling image features and another for handling word sequence pair. Firstly, image features were extracted using InceptionResnetV2 [28] or Xception [38]. These image features were preceded down to a dropout layer followed by a fully connected layer and then another dropout layer. A fully connected layer was used to reduce the dimension of the image feature vector from 1536 or 2048 to 256 to match the dimension of word prediction output. Secondly, Input word sequence pairs are feed to two embedding layers one was pre-trained GloVe embedding and another was fastText which was not pre-trained. Both embeddings were used to convert words to vectors of dimension 100. The vector from the two embeddings was then passed through a separate dropout layer followed by either BiLSTM or BiGRU of dimension 128. To match the dimension of visual feature vector output these vectors were passed through an additional fully connected layer of dimension 256. These two outputs were then concatenated. This concatenated output was then mapped to the visual part of the encoder using another concatenation and then forwarded to the decoder.

- Decoder
  The decoder is a Feed Forward Network which ends with a SoftMax. It takes the concatenated output of the encoder as input. This input was first passed through a fully connected layer of 256 dimensions followed by a dropout layer. Finally, via probabilistic Softmax function outputs the next word in the sequence. The SoftMax greedily selects the word with maximum probability.

IV. Experimental Setup

This section narrates the total strategy adapted to obtain captions from images. Also, different tuning techniques availed are described here.

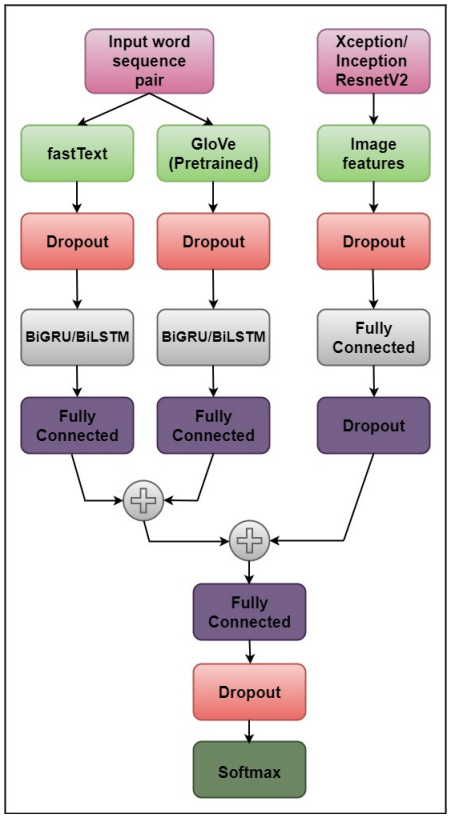


Fig. 3. Proposed Hybrid Model.

A. Dataset Processing

Flickr8k dataset has 8091 images of which 6000 (75%) images are employed for training, 1000 (12.5%) images for validation and 1000 (12.5%) images are used for testing. Moreover, with each image of the Flickr8K dataset five ground truth captions describing the image are designated which adds up to a total of 40455 captions for 8091 images. For image captioning in Bengali, those 40455 captions were converted

to Bengali language using Google Translator. Unfortunately, some of the translated captions were syntactically incorrect. Hence, we manually checked all 40455 translated captions and corrected them. We utilized these 8000 images as well as selected 4000 images as done by Deb et al. [5] in Bengali(Flickr4k-BN and Flickr8k-BN). These 4000 images were selected based on the frequency of words in those 40455 captions. Using POS taggers most frequent nouns Bengali words were identified from ground truth captions. The most frequent words in the Bengali Flickr8k dataset are shown in Fig. 4 for Bengali and English respectively. 4000 images analogous to these words are selected and made two small datasets Flickr4k-BN.

We also utilized the BanglaLekha dataset which consists of 9154 images. It is the only available Bengali dataset till now. All its captions are human annotated. One problem with this dataset is that it has only two captions associated with each image resulting in 18308 captions for those 9154 images. Hence, vocabulary size is lower than Flickr4k-BN and Flickr8k-BN. Flickr8k-BN consists of 12953 unique Bengali words, Flickr4k-BN consist of 6420 unique Bengali words and BanglaLekha consists of 5270 unique Bengali words. It can be seen that the BanglaLekha dataset has a vocabulary size even lower than Flickr4k-BN. Hence, we employed the Flickr8k-BN dataset alongside Flickr4k-BN and BanglaLekha datasets. The split ratio of all three datasets for training, testing and validating are shown in Table I.

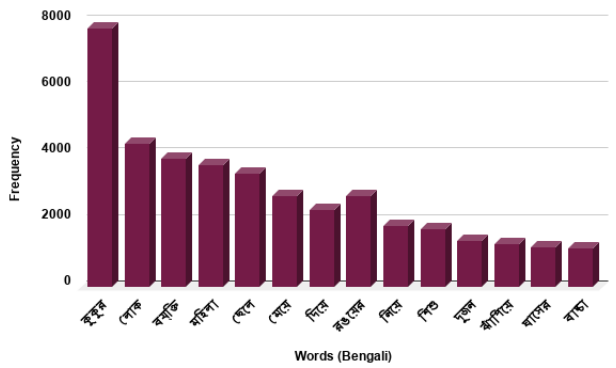


Fig. 4. Illustration of Most Frequent Noun Bengali Words in Flickr8k Bengali Dataset.

B. Image Feature Extraction

One essential part of image captioning is to extract features from given images. This task is achieved using Convolutional Neural Network architectures. These architectures are used to detect objects from images. They can be trained on a large number of images for extracting image features. This training process requires an enormous number of images and time. Due to the shortage of a large number of images, we utilized Convolutional Neural Network architecture which was pre-trained on more than a million images from the ImageNet [33] dataset in our model known as InceptionResnetV2 [28] and Xception [38]. These two pre-trained architectures were used separately for different experimental setups. The reason for using InceptionResnetV2 and Xception is that these models can achieve higher accuracy at lower epochs. The last layer which is used for prediction purposes of this pre-trained of InceptionResnetV2 model is pulled out and the last two layers of the pre-trained Xception model were pulled out. Finally, the average pooling layer was used to extract image features and convert them into a feature vector of 1536 dimensions for InceptionResnetV2 and 2048 dimensions for Xception. All the images are given an input shape of 299x299x3 before entering the InceptionResnetV2 model. Here 3 represents the three-color channels R, G and B.

C. Embeddings

Handling word sequences requires word embedding that can convert words to vectors before passing them to special recurrent neural networks (RNN). In our model GloVe [22] and fastText [7] have been used as an embedding.

- GloVe is a model for distributed word representation. The model employs an unsupervised learning algorithm for acquiring vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity.
- fastText is a library for the learning of word embeddings and text classification created by Facebook's AI Research (FAIR) lab. The model employs unsupervised learning or supervised learning algorithms for obtaining vector representations for words. fastText yields two models for computing word representations namely skipgram and cbow. Skipgram model learns to forecast a target word using the nearby word. conversely, cbow model forecasts the target word according to its context where context depicts a bag of the words contained in a fixed size window around the target word.

Both GloVe and fastText have pre-trained word vectors that are trained over a large vocabulary. These embeddings can also be trained. In the hybrid model shown in Fig. 3, two embeddings have been used GloVe and fastText. There GloVe was pre-trained but fastText has been trained on vocabulary available in the dataset. Trainable fastText instead of pre-trained fastText was used to enrich the vocabulary with words in Flickr8k and BanglaLekha datasets. Also, results of pre-trained fastText have already been demonstrated by Deb et al. [5]. The combination of two embedding leads to redundancy of words but it gives fluent caption in Bengali as the vocabulary size increases. On the other hand, pre-trained files for both GloVe and fastText in the hybrid model will give much greater redundancy and the vocabulary size becomes small as the vocabulary does not contain unique words in the dataset.

Two other models were trained alongside the hybrid model. Unlike the hybrid model, these two models had a single embedding either a trainable fastText embedding or a pre-trained GloVe embedding. GloVe file "bn_glove.39M.100d"[4] pre-trained in Bangali Language was used for Bengali datasets.

D. Word Sequence Generation

Flickr8k dataset has five captions associated with each image and BanglaLekha has two captions associated with each

[4]https://github.com/sagorbrur/GloVe-Bengali

TABLE I. DISTRIBUTION OF DATA FOR THREE BENGALI DATASET USED. SAME DISTRIBUTION WAS USED FOR FLICKR8K ENGLISH AND BENGALI DATASETS.

Dataset	Total Image	Training	Validation	Testing
Flickr4k	4000	2400 (60%)	800 (20%)	800 (20%)
Flickr8k	8000	6000 (75%)	1000 (15%)	1000 (15%)
BanglaLekha	9154	7154 (78%)	1000 (11%)	1000 (11%)

image. One of the difficult tasks of image captioning is to make the model learn how to generate these sentences. Two different types of special Recurrent Neural Network (RNN) were used to train the model to generate the next word in the sequence of a caption. The input and output sizes were fixed to the maximum length of the sentence present in the dataset. In the case of Flickr4k-BN and Flickr8k-BN maximum length was 23. On the other hand, two different maximum lengths of the sequence 40 and 26 were used for the BanglaLekha dataset. Reducing the maximum sequence length significantly increased the evaluation scores. While training if any sentence were generated having a length less than the maximum length zero-padding was applied to make that sentence length equal to the fixed length. Additionally, an extra start token and end token is added to the sequence pair for identification in the training process. During training, image features vector and previous words converted to vector using embedding layer were used to generate the next word in the sequence probabilistic Softmax with the help of different types of RNN. Fig. 5 illustrates the input and output pair.

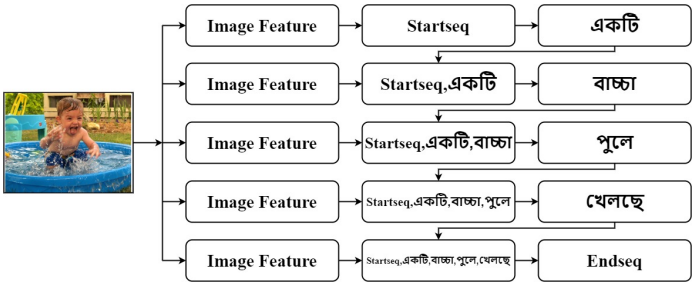


Fig. 5. Demonstrates How Word Sequences are Generated.

Due to the limitation of the basic Recurrent Neural Network (RNN) [34] to retrain long term memory a better approach was taken by Deb et al. [5] which uses Long Short-Term Memory (LSTM). However, LSTM [10] only preserve preceding words but for proper sentence generation succeeding words are also necessary. As a result, our model uses Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) which are illustrated in Fig. 6. Each box marked as A or A' was either a Long Short-Term Memory (LSTM) or a Gated Recurrent Unit GRU [8] unit. X [0...i] are the input words and Y [0...i] are the output words. Y [0...i] are determined using the Eq. 1.

$$\hat{y}^{<t>} = g(W_y[\overrightarrow{a}^{<t>}\overleftarrow{a}^{<t>}] + b_y) \tag{1}$$

Where $\hat{y}^{<t>}$ is the output at time t when activation function g is applied to recurrent component's weight W_y and bias by with both forward activation $\overrightarrow{a}^{<t>}$ at time t and backward activation $\overleftarrow{a}^{<t>}$ at time t.

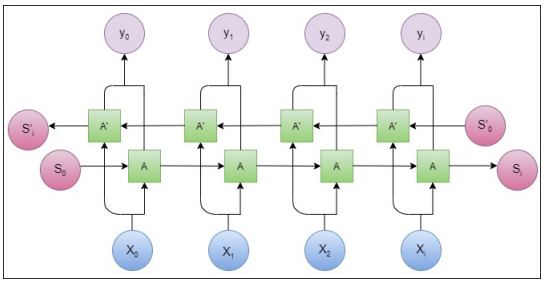


Fig. 6. Illustrates Bidirectional RNN having $X_0 \ldots i$ as Input and $Y_0 \ldots i$ as output. A and A' boxes are both either BiLSTM or BiGRU where A is the Forward Recurrent Component an A' is the Backward Recurrent Component.

- GRU is a special type of RNN. Reset and update the gate of a GRU helps to solve the vanishing gradient problem of RNN. The update gate of GRU seeks how much information from the previous units must be forwarded. The update gate adopted is computed by the following formula:

$$z_t = \sigma(W_z.[h_{t-1}, x_t]) \tag{2}$$

where z_t is update gate output at the current timestamp, W_z is weight matrix at update gate, h_{t-1} information from previous units, and x_t is input at the current unit.
The reset gate is used by the model to find how much information from the previous units to forget. The reset gate is computed by the following formula:

$$r_t = \sigma(W_r.[h_{t-1}, x_t]) \tag{3}$$

where r_t is reset gate output at current timestamp, W_r is weight matrix at reset gate, h_{t-1} information from previous units, and x_t is input at the current unit.
Current memory content used to store the relevant information from the previous units. It is calculated as follows:

$$\tilde{h}_t = tanh(W.[r_t * h_{t-1}, x_t]) \tag{4}$$

where $\tilde{h}_t$ is current memory content, W is weight at current unit, r_t is reset gate output at current timestamp, h_{t-1} is information from previous units, and x_t is input at the current unit.
Final memory at the current unit is a vector used to store the final information for the current unit and pass it to the next layer. It is calculated using a formula:

$$h_t = (1 - z_t) * h_{t-1} + z_t\tilde{h}_t \tag{5}$$

where h_t is final memory at the current unit, z_t is update gate output at current timestamp, h_{t-1} is information from previous units, and $\tilde{h}_t$ is current memory content.

- LSTM is another Special type of RMNN. Unlike the GRU the LSTM has three gates, namely, the forget gate, update gate and the output gate. The equations for the gates in LSTM are:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{6}$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{7}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{8}$$

where i_t represents input gate, f_t represents forget gate, o_t represents output gate, σ represents sigmoid function, W_x represents weight of the respective gate(x) neurons, h_{t-1} represents output of previous LSTM block at timestamp t-1, x_t represents input at current timestamp and b_x represents biases for the respective gates(x).
Input gate tells what new information is going to be stored in cell state. Forget gate determine what information to throw away from cell state and Output get is used to provide output at timestamp t. The equations for the cell state, candidate cell state and the final output are:

$$\tilde{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \tag{9}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{10}$$

$$h_t = o_t * tanh(c^t) \tag{11}$$

where c_t represents cell state at timestamp t and $\tilde{c}_t$ represent candidate for cell state at timestamp t. candidate timestamp must be generated to get memory vector for current timestamp c_t. Then the cell state is passed through a activation function to generate h_t. Finally, h_t is passed through a softMax layer to get the output y_t.

E. Hyperparameter Selection

One major problem of machine learning is overfitting. Overfit models have high variance. These models cannot generalize well. As a result, this is a huge problem for image captioning. We observed the performance of our model and noticed that it was suffering from overfitting rather than underfitting. To minimize this overfitting problem some hyperparameter tuning has been adapted in our model. Firstly, different values of dropout [35] have been used for sequence model image features and decoder. Dropouts help prevent overfitting. For feature extractor dropout value of 0.0 was used, a dropout of 0.3 was used for the sequence model and in the case of decoder dropout value of 0.5 was utilized. Secondly, different activation functions were employed for different fully connected layers. For example, regarding the feature extractor model and decoder ELU [3] activation function was availed and for the sequence model, ReLU [36] activation function was employed. Thirdly, we employed external validation to provide an unbiased evaluation and ModelCheckpoint was availed to save models that had minimum validation loss. On the other hand, ReduceLROnPlateau was used for models that had Xception as CNN. Moreover, Adam optimizer [14] was utilized and the models were trained for 50 and 100 epochs having learning rates of 0.0001 and 0.00001. A short summary of the hyperparameters adapted in different models are shown in Table II and the loss plot of BanglaLekha dataset and Flickr8K-BN dataset are ornamented in Fig. 7 and Fig. 8, respectively. From these plots, it can be seen that the model converges towards epoch 100. Another important factor that improved the result was maximum sentence length. In the BnglaLekha only a few sentences had lengths greater than 26. As a result, we took a maximum length of sentences in this dataset to 26. This enhanced the evaluation scores greatly.

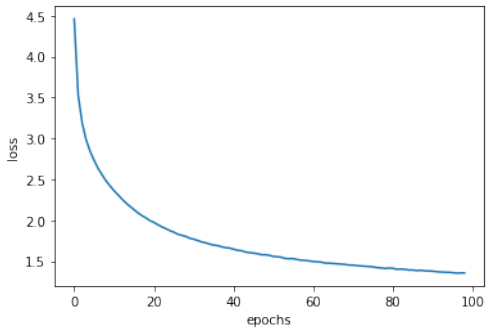


Fig. 7. Loss Plot of BanglaLekha Dataset for 100 epochs.

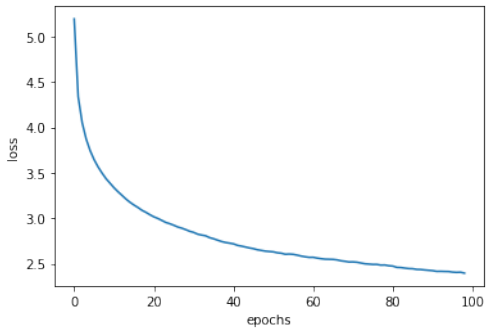


Fig. 8. Loss Plot of Flickr8k-BN Dataset for 100 epochs.

V. Analysis

We implemented the algorithm using Keras 2.3.1 and Python 3.8.1. Additionally, we ran our experiments on GPU RTX 2060. Our code and Bengali Flickr8k dataset is given in GitHub[5]. We translated the Flickr8k dataset to Bengali using Google Translator Like that done by [16]. Bilingual Evaluation Understudy (BLEU) [20] score was used to evaluate the performance of our models as it is the most wielded metric nowadays to evaluate the caliber of text. It depicts how normal sentences are compared with human generated sentences. It is broadly utilized to evaluate the performance of Machine translation. Sentences are compared based on modified n-gram precision for generating BLEU scores. BLEU scores are computed using the following equations:

$$P(i) = \frac{Matched(i)}{H(i)} \tag{12}$$

[5]https://github.com/MayeeshaHumaira/A-Hybridized-Deep-Learning-Method-for-Benga

TABLE II. HYPERPARAMETERS ADAPTED IN DIFFERENT MODELS.

Search Type	Model	Learning Rate	Loss Function	Callback	Epoch
Greedy	Xception +BiLSTM	0.00001	Sparse Categorical Crossentrophy	ReduceLROnPlateau	100
	InceptionResnetV2 +BiLSTM	0.0001	Categorical Crossentrophy	ModelCheckpoint	50
Beam=3	Xception +BiLSTM	0.00001	Sparse Categorical Crossentrophy	ReduceLROnPlateau	100
Beam=5	Xception +BiLSTM	0.00001	Sparse Categorical Crossentrophy	ReduceLROnPlateau	100

where P(i) is the precision that is for each i-gram where i = 1, 2, ...N, the percentage of the i-gram tuples in the hypothesis that also occur in the references is computed. H(i) is the number of i-gram tuples in the hypothesis and Matched(i) is computed using the following formula:

$$Matched(i) = \sum_{t_i} \min\{C_h(t_i), \max_j C_{hj}(t_i)\} \quad (13)$$

where t_i is an i-gram tuple in hypothesis h, $C_h(t_i)$ is the number of times t_i occurs in the hypothesis, $C_{hj}(t_i)$ is the number of times t_i occurs in reference j of this hypothesis.

$$\rho = exp\{\min(0, \frac{n-L}{n})\} \quad (14)$$

where ρ is brevity penalty to penalize short translation, n is the length of the hypothesis and L is the length of the reference. Finally, the BLEU score is computed by:

$$BLEU = \rho\{\prod_{i=1}^{N} P(i)\}^{\frac{1}{N}} \quad (15)$$

Two different search types Greedy and Beam search were used to compute these BLEU scores. In a Greedy search word with maximum probability is chosen as the next word in the sequence. On the other hand, Beam search considers n words to choose from for the next word in the sequence. Where n is the width of the beam. For our experiment, we considered beamwidth of 3 and 5. We computed 1-gram BLEU (BLEU-1), 2-gram BLEU (BLEU-2), 3-gram BLEU (BLEU-3), 4-gram BLEU (BLEU-4) for various architectures. These are illustrated in Table III, Table IV and Table V.

Performance of the proposed Hybrid architecture and single embedding GloVe or fastText on Flickr4k-BN dataset consisting of 4000 data for Bengali are demonstrated in Table III. From Table III it can be stated that the Hybrid model performed better for both BiLSTM and BiGRU on the Bengali dataset than only GloVe and only fastText word embedding. Moreover, we obtained better BLEU scores than paper [5]. The greedy search was employed to compute these BLEU scores.

Consequently, the performance of the single embedding GloVe or fast Text and hybrid architecture on Flickr8k-BN dataset consisting of 8000 data and BanglaLekha dataset are displayed in Table IV and Table V, respectively. There also it can be observed that the proposed Hybrid model performed better for both BiGRU and BiLSTM than the other models. The Highest BLEU score was obtained using BiLSTM on Flickr4k-BN and Flickr8k-BN as a result the captions generated by the Hybrid model for both datasets are illustrated in Fig. 9. Furthermore, our proposed Hybrid model also gave the highest BLEU scores for the BanglaLekha dataset for both BiLSTM and BiGRU as shown in Table V. From there it can be observed that Xception and the learning rate played a vital role in increasing the BLEU scores. These scores were even better than BLEU scores obtained by paper [37]. Table VI illustrates a brief comparison of the BLEU scores obtained by our proposed model and the scores obtained by other papers. From there it can be observed that our proposed Hybrid model indeed gave a better performance. The captions generated by these models for test images of the BanglaLekha dataset are shown in Fig. 10. Flickr8k-BN dataset consisting of 8000 images were not previously used by any other papers for generating captions in Bengali.

Fig. 9. Illustration of Captions Generated by Best Performing Hybrid Architecture using Fickr4k-BN and Flickr8k-BN Datasets.

TABLE III. RESULT OF INCEPTIONRESNETV2 USED BY FLICKR4K-BN

Experimental Model	RNN	Training Accuracy	Validation Accuracy	BLEU			
				1	2	3	4
Proposed	BiLSTM	0.421	0.387	**0.661**	**0.508**	**0.382**	**0.229**
Hybrid architecture	BiGRU	0.432	0.386	0.660	0.503	0.371	0.215
GloVe	BiLSTM	0.432	0.388	0.644	0.491	0.369	0.220
	BiGRU	0.429	0.386	0.651	0.497	0.373	0.223
fastText	BiLSTM	0.414	0.372	0.638	0.490	0.370	0.219
	BiGRU	0.426	0.379	0.653	0.505	0.381	0.226

TABLE IV. BLEU SCORES OBTAINED USING FLICKR8K-BN DATASET

Search Type	Learning Rate	Word Embedding	Experimental Model	BLEU			
				1	2	3	4
Greedy	0.00001	Hybrid	Xception+BiLSTM	0.504	0.326	0.232	0.119
			Xception+BiGRU	0.536	0.352	0.246	0.126
		GloVe	Xception+BiLSTM	0.539	0.356	0.249	0.129
			Xception+BiGRU	0.532	0.352	0.241	0.121
		fastText	Xception+BiLSTM	0.190	0.055	0.000	0.000
			Xception+BiGRU	0.194	0.068	0.012	0.000
Greedy	0.0001	Hybrid	InceptionResnetV2+BiLSTM	**0.540**	**0.370**	**0.268**	**0.145**
			InceptionResnetV2+BiGRU	0.526	0.360	0.261	0.141
		GloVe	InceptionResnetV2+BiLSTM	0.534	0.369	0.265	0.142
			InceptionResnetV2+BiGRU	0.512	0.350	0.255	0.138
		fastText	InceptionResnetV2+BiLSTM	0.528	0.363	0.269	0.140
			InceptionResnetV2+BiGRU	0.530	0.362	0.260	0.140
Beam=3	0.00001	Hybrid	Xception+BiLSTM	0.416	0.246	0.176	0.089
			Xception+BiGRU	0.414	0.247	0.178	0.093
		GloVe	Xception+BiLSTM	0.395	0.239	0.174	0.089
			Xception+BiGRU	0.404	0.245	0.178	0.090
		fastText	Xception+BiLSTM	0.034	0.000	0.000	0.000
			Xception+BiGRU	0.059	0.003	0.001	0.000
Beam=5	0.00001	Hybrid	Xception+BiLSTM	0.409	0.240	0.175	0.090
			Xception+BiGRU	0.403	0.239	0.171	0.089
		GloVe	Xception+BiLSTM	0.377	0.226	0.162	0.079
			Xception+BiGRU	0.393	0.241	0.172	0.085
		fastText	Xception+BiLSTM	0.034	0.000	0.000	0.000
			Xception+BiGRU	0.059	0.003	0.001	0.000

VI. CONCLUSION

In this work, we exhibited a notion for automatically generating caption from an input image in Bengali. Firstly, a detailed description of how the Flickr8k dataset was translated in Bengali and distributed into a dataset of two sizes was presented. Secondly, how image features were extracted and the different combinations of word embedding utilized were also conferred. Moreover, the reasons for using a special kind of word sequence generator was elucidated. Furthermore, different parts of the proposed architecture were ornamented. Finally, using the BLEU score it was authenticated that the proposed architecture performs better for both Flickr4k-Bn and BanglaLekha datasets. This validates the fact that image captioning using the Bengali language can be refined further in the future. We will try to adapt the visual attention and transformer model in the near future for better feature extraction and getting more precise captions. Additionally, we aim to make our own dataset having five captions with each image, unlike the BanglaLekha dataset that has two captions associated with each image to enrich the vocabulary of our dataset.

REFERENCES

[1] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 2, pp. 417–421, 2017, doi: 10.18653/v1/P17-2066.

[2] J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study of Language CNN for Image Captioning," Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-October, pp. 1231–1240, 2017, doi: 10.1109/ICCV.2017.138.

[3] D. A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc., pp. 1–14, 2016.

[4] W. Cui et al., "Landslide image captioning method based on semantic gate and bi-temporal LSTM," ISPRS Int. J. Geo-Information, vol. 9, no. 4, 2020, doi: 10.3390/ijgi9040194.

[5] T. Deb et al., "Oboyob: A sequential-semantic Bengali image captioning engine," J. Intell. Fuzzy Syst., vol. 37, no. 6, pp. 7427–7439, 2019, doi: 10.3233/JIFS-179351.

[6] H. Dong, J. Zhang, D. Mcilwraith, and Y. Guo, "I2T2I: LEARNING TEXT TO IMAGE SYNTHESIS WITH TEXTUAL DATA AUGMENTATION."

[7] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval., pp. 3483–3487, 2019.

[8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," pp. 1–9, 2014, [Online]. Available: http://arxiv.org/abs/1412.3555.

[9] M. Han, W. Chen, and A. D. Moges, "Fast image captioning using LSTM," Cluster Comput., vol. 22, pp. 6143–6155, May 2019, doi: 10.1007/s10586-018-1885-9.

TABLE V. BLEU SCORES OBTAINED USING BANGLALEKHA DATASET

Search Type	Learning Rate	Word Embedding	Experimental Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4
Greedy	0.00001	Hybrid	Xception+BiLSTM	0.673	0.525	0.454	0.339
			Xception+BiGRU	**0.674**	**0.527**	**0.454**	**0.344**
		GloVe	Xception+BiLSTM	0.612	0.453	0.380	0.265
			Xception+BiGRU	0.610	0.454	0.383	0.272
		fastText	Xception+BiLSTM	0.618	0.463	0.389	0.277
			Xception+BiGRU	0.624	0.473	0.402	0.290
Greedy	0.0001	Hybrid	InceptionResnetV2+BiLSTM	0.568	0.396	0.287	0.160
			InceptionResnetV2+BiGRU	0.571	0.402	0.301	0.173
		GloVe	InceptionResnetV2+BiLSTM	0.568	0.401	0.301	0.174
			InceptionResnetV2+BiGRU	0.570	0.403	0.303	0.176
		fastText	InceptionResnetV2+BiLSTM	0.553	0.390	0.291	0.169
			InceptionResnetV2+BiGRU	0.567	0.398	0.300	0.171
Beam=3	0.00001	Hybrid	Xception+BiLSTM	0.434	0.344	0.303	0.234
			Xception+BiGRU	0.411	0.324	0.286	0.221
		GloVe	Xception+BiLSTM	0.383	0.285	0.245	0.176
			Xception+BiGRU	0.401	0.302	0.263	0.196
		fastText	Xception+BiLSTM	0.419	0.320	0.283	0.214
			Xception+BiGRU	0.434	0.329	0.293	0.221
Beam=5	0.00001	Hybrid	Xception+BiLSTM	0.420	0.335	0.297	0.232
			Xception+BiGRU	0.399	0.316	0.280	0.219
		GloVe	Xception+BiLSTM	0.368	0.273	0.234	0.170
			Xception+BiGRU	0.385	0.292	0.256	0.194
		fastText	Xception+BiLSTM	0.422	0.324	0.288	0.219
			Xception+BiGRU	0.429	0.326	0.291	0.222

TABLE VI. A BRIEF COMPARISON OF BLEU SCORES FOR EXISTING MODELS AND OUR PROPOSED HYBRID MODEL.

Dataset	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4
BanglaLekha	VGG-16+LSTM [37]	66.7	43.6	31.5	23.8
	Xception+BiGRU (Our Hybrid Model)	**0.674**	**0.527**	**0.454**	**0.344**
Flickr8k(4000 images)	Merge Bengali(Inception+LSTM) [5]	0.62	0.45	0.33	0.22
Flickr4k-BN	Our Hybrid Model (InceptionResnetV2+BiLSTM)	**0.661**	**0.508**	**0.382**	**0.229**

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[11] K. Xu, H. Wang, and P. Tang, "IMAGE CAPTIONING WITH DEEP LSTM BASED ON SEQUENTIAL RESIDUAL Department of Computer Science and Technology , Tongji University , Shanghai , P . R . China Key Laboratory of Embedded System and Service Computing , Ministry of Education ," no. July, pp. 361–366, 2017.

[12] V. Jindal, "Generating Image Captions in Arabic Using Root-Word Based Recurrent Neural Networks and Deep Neural Networks." Available: www.aaai.org.

[13] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," 2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019, pp. 1–6, 2019, doi: 10.1109/ICoICT.2019.8835370.

[14] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

[15] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," MM 2017 - Proc. 2017 ACM Multimed. Conf., pp. 1549–1557, 2017, doi: 10.1145/3123266.3123366.

[16] X. Li, W. Lan, J. Dong, and H. Liu, "Adding Chinese captions to images," ICMR 2016 - Proc. 2016 ACM Int. Conf. Multimed. Retr., pp. 271–275, 2016, doi: 10.1145/2911996.2912049.

[17] C. Liu, F. Sun, and C. Wang, "MMT: A multimodal translator for image captioning," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10614 LNCS, p. 784, 2017.

[18] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks," pp. 1–9, 2014, [Online]. Available: http://arxiv.org/abs/1410.1090.

[19] Q. You, H. Jin, Z. Wang, ... C. F.-P. of the I., and undefined 2016, "Image captioning with semantic attention," openaccess.thecvf.com Available: http://openaccess.thecvf.com/.

[20] K. Papineni, S. Roukos, T. Ward, W. Zhu, and Y. Heights, "IBM Research Report Bleu: a Method for Automatic Evaluation of Machine Translation," Science (80-.)., vol. 22176, no. February, pp. 1–10, 2001, doi: 10.3115/1073083.1073135.

[21] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.

[22] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., no. June, pp. 1532–1543, 2014, doi: 10.3115/v1/d14-1162.

[23] M. Rahman, N. Mohammed, N. Mansoor, and S. Momen, "Chittron: An Automatic Bangla Image Captioning System," Procedia Comput. Sci., vol. 154, pp. 636–642, 2018, doi: 10.1016/j.procs.2019.06.100.

[24] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image Captioning Based on Deep Neural Networks," MATEC Web Conf., vol. 232, pp. 1–7, 2018, doi: 10.1051/matecconf/201823201052.

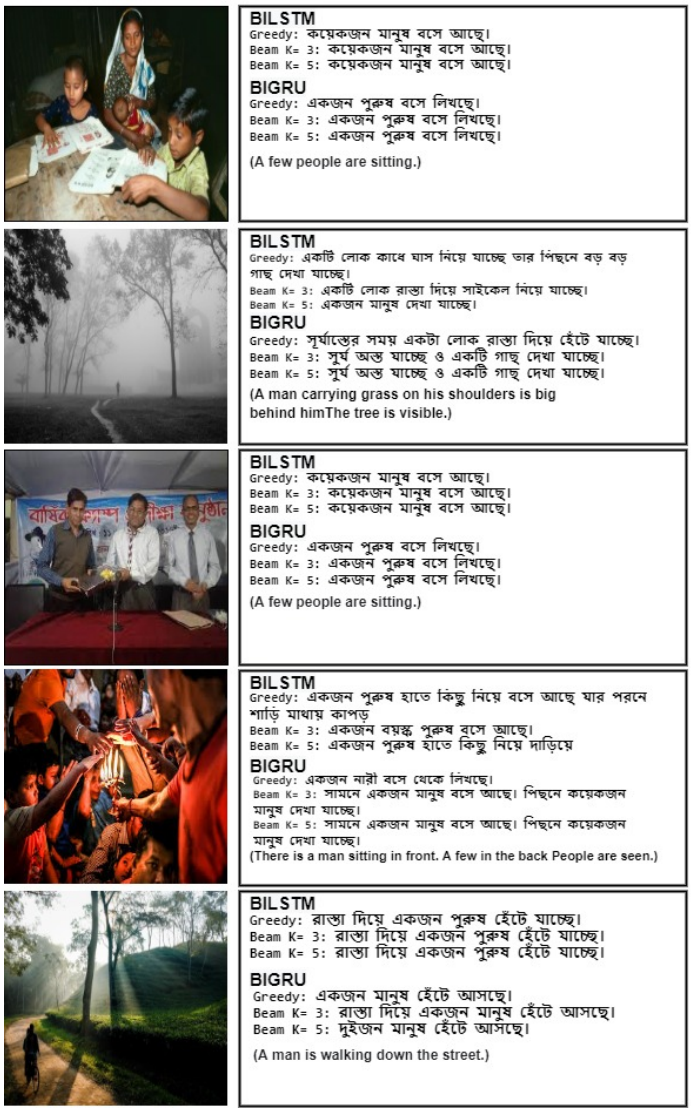


Fig. 10. Illustration of Captions Generated by Best Performing Hybrid Architecture using BanglaLekha Dataset.

[25] [1] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." Available: http://proceedings.mlr.press/v37/xuc15.

[26] S. R. Laskar, R. P. Singh, P. Pakray, and S. Bandyopadhyay, "English to Hindi Multi-modal Neural Machine Translation and Hindi Image Captioning," pp. 62–67, 2019, doi: 10.18653/v1/d19-5205.

[27] R. Subash, R. Jebakumar, Y. Kamdar, and N. Bhatt, "Automatic image captioning using convolution neural networks and LSTM," J. Phys. Conf. Ser., vol. 1362, no. 1, 2019, doi: 10.1088/1742- 6596/1362/1/012096.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 4278–4284, 2017.

[29] A. Jaffe, "Generating Image Descriptions using Multilingual Data," pp. 458–464, 2018, doi: 10.18653/v1/w17-4750.

[30] C. Wang, H. Yang, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning," ACM Trans. Multimed. Comput. Commun. Appl., vol. 14, no. 2s, 2018, doi: 10.1145/3115432.

[31] Y. Xian and Y. Tian, "Self-Guiding Multimodal LSTM - When We Do Not Have a Perfect Training Dataset for Image Captioning," IEEE Trans. Image Process., vol. 28, no. 11, pp. 5241–5252, 2019, doi: 10.1109/TIP.2019.2917229.

[32] P. Blandfort, T. Karayil, D. Borth, and A. Dengel, "Image captioning in the wild: How people caption images on flickr," MUSA2 2017 - Proc. Work. Multimodal Underst. Soc. Affect. Subj. Attrib. co-located with MM 2017, pp. 21–29, 2017, doi: 10.1145/3132515.3132522.

[33] L. Fei-Fei, J. Deng, and K. Li, "ImageNet: Constructing a large-scale image database," J. Vis., vol. 9, no. 8, pp. 1037–1037, 2010, doi: 10.1167/9.8.1037.

[34] M. Tanti, A. Gatt, and K. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?," pp. 51–60, 2018, doi: 10.18653/v1/w17-3506.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," 2014. doi: 10.5555/2627435.2670313.

[36] [1] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines." Available: https://www.cs.toronto.edu/ hinton/absps/reluICML.pdf.

[37] A. H. Kamal, M. A. Jishan and N. Mansoor, "TextMage: The Automated Bangla Caption Generator Based On Deep Learning," 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2020, pp. 822-826, doi: 10.1109/DASA51403.2020.9317108.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 1800–1807, 2017, doi: 10.1109/CVPR.2017.195.

Hybrid Approaches based on Simulated Annealing, Tabu Search and Ant Colony Optimization for Solving the k-Minimum Spanning Tree Problem

El Houcine Addou[1], Abelhafid Serghini[2]
LANO Laboratory FSO-ESTO Mohammed I University
BV Mohamed VI BP 717, Oujda 60000, Morocco

El Bekkaye Mermri[3]
FSO Mohammed I University
BV Mohamed VI BP 717, Oujda 60000, Morocco

***Abstract*—In graph theory, the k-minimum spanning tree problem is considered to be one of the well-known NP hard problems to solve. This paper address this problem by proposing several hybrid approximate approaches based on the combination of simulated annealing, tabu search and ant colony optimization algorithms. The performances of the proposed methods are compared to other approaches from the literature using the same well-known library of benchmark instances.**

***Keywords*—k-Minimum spanning tree; metaheuristics; simulated annealing; ant colony optimization algorithms; tabu search; approximation algorithms**

I. Introduction

In this work we attempt to provide some approximate methods to solve the well-known combinatorial optimization: The k-minimum spanning tree (k-MST) problem. We want to find a tree in an edge-weighted graph $G = (V, E)$ that have exactly k edges and which minimize the sum of the weights of its edges. The mathematical formulation of this problem has been introduced by [1], It was demonstrated that the k-MST problem is known to be NP-hard [2], it is very difficult to find optimal solutions to large-scale problems that can be formulated as a k-MST within acceptable time.

In literature, several approaches using metaheuristics were proposed to tackle the k-MST problem [3], [4], [6], [5], [7]. In [8], three approaches to deal with the k-MST were presented: an evolutionary computation, ant colony optimization (ACO) and tabu search (TS), in order to show and compare the performance of these methods, the authors built a library named KCTLIB which contains some graph instances for k-MST problems, after that they executed their programs, numerical results showed that ACO algorithm is the best choice to deal with k-MST with small cardinalities, while TS is the best choice for k-MST with large cardinalities. these conclusions had inspired the authors in [14] to hybridize TS with ACO and the authors in [16] to hybridize TS with SA, they have applied their approaches to some graph instances from KCTLIB, and they have presented the results of their methods.

The objective of this paper is to attempt new hybrid approaches by coupling simulated annealing (SA), TS and ACO algorithms. We aim to find new best k-MST solutions, numerical experiments were performed using two graph instances from KCTLIB. Our experiments show that the suggested approaches are able to find new best values for the two graphs and for several cardinalities.

The paper is structured as follows. The problem formulation is presented in section II . In the section III we give the description of the main components of SA, TS, and ACO algorithms. the results of the computational experiments and the discussion are reported in section IV . Finally, in section V we give some conclusions.

II. Problem Formulation

Given an undirected edge-weighted graph $G = (V, E)$ on a set V of n vertices , and a positive cost function $w(e)$ on the set of edges E, $|V|$ is the number of vertices of G, $|E|$ is the number of edges. A spanning tree (ST) of G is a connected subgraph that contains all vertices and without any cycle. The k-spanning tree ($k \leq |V| - 1$) that we note T_k is a tree that contains k edges, if $k = |V| - 1$ we get a spanning tree. we note by X_k the set of possible k-spanning trees, The set of all possible k-spanning trees is denoted by X_k. the k-minimum spanning tree (k-MST) problem asks for a k-spanning tree with the minimum sum of weights. The k-MST problem can be formulated as follows:

$$\begin{cases} \text{Minimize} & \displaystyle\sum_{e \in E(T_k)} w(e) \\ \text{Subject to} & T_k \in X_k, \end{cases}$$

$E(T_k)$ denotes edges set of T_k .

An optimal solution can be easily found in case of small problem size by enumerating all k-spanning trees in a given graph. However, it has been demonstrated that the k-MST problem is a NP-hard problem. Therefore, it is very important to develop new approximate methods using metaheuristics in order to provide solutions to real-world problems in reasonable time.

III. Proposed Approaches

A. Simulated Annealing

SA is an approximate algorithm inspired from thermodynamics, which was introduced in [9] and it is widely used to address many discrete and continuous optimization problems, such as the travelling salesman, and vehicle routing problem, etc. Researchers are also applying it to multi-criteria

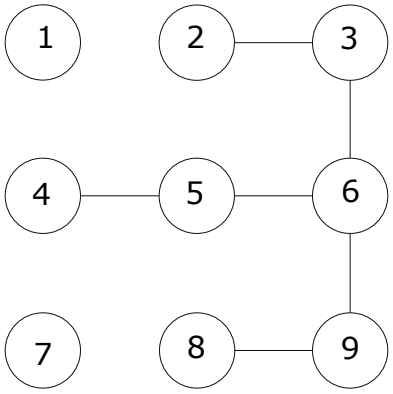


Fig. 1. A 6-ST of 7 Vertices.

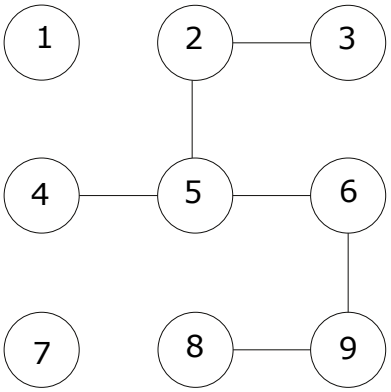


Fig. 2. A Neighborhood Element of the 6-ST.

optimization problems [10], [11]. SA is a neighborhood search method. The basic concept of this method is to move at each iteration from a solution x to another one that belongs to its neighborhood $N(x)$. Before describing the main SA components, we present the neighborhood structure used in this method.

B. The Neighborhood

To move around and allow the algorithm to discover the search space, the possible transitions or movements from one solution to another are subject to the chosen neighborhood structure. The neighborhood of a tree T_k is the set of k-ST built by deleting one edge from T_k and replacing it by an edge of the graph G and which doesn't belong to T_k. We choose randomly an edge from the current T_k and replace it by another one, which is also is selected randomly from $G \backslash T_k$. To clarify the neighborhood structure, we take the example of a tree composed by six edges in Fig. 1. The 6-ST contains the following edges: (2,3),(3,6),(6,5),(5,4),(6,9),(9,8).

Fig. 2 represents a neighborhood element of the 6-ST in Fig. 1, where the edge (3,6) is replaced by the edge (2,5).

C. SA Algorithm

The SA settings are adjusted empirically, as follows, the main lines of the proposed SA algorithm are as follows:

1) Set the initial values of the following parameters:
   a) T_0: The initial temperature
   b) α: Factor of cooling
   c) TMP_LEVELS: The maximum number of temperature levels; the temperature is not decreased at each iteration but it is decreased by level
   d) TMP_RANGE: Number of iterations per level
   e) $TIME\_TO\_DIVERSIFY$: the time limit of the algorithm to launch the diversification based on the ACO.
2) Initial solution: Use the Prim method to build a k-subtree
3) SA procedure
   - Until TMP_LEVELS repeat
   a) Repeat for TMP_RANGE iterations:
      - Select randomly a k-ST in the neighborhood of the current solution.
      - Calculate the objective function of the current solution f_1
      - Calculate the objective function of the neighborhood solution f_2
      - Calculate F = f_2 - f_1
        If $F < 0$ then set the neighborhood element as the current solution
        Otherwise, the neighborhood element will be the current solution with a probability equal to $exp(-F/T)$, T is the value of the temperature in the current iteration
   b) Use the geometric cooling schedule to decrease the current temperature

D. Tabu Search

TS was introduced by F. Glover and Laguna in [12], [17], [13]. This metaheuristic method is used at large scale to find approximate solutions for real-world optimization problems, TS can be applied for example in logistics, resource planning, telecommunications, scheduling, etc.

TS is based on simple ideas inspired from the human memory, it is a local search method which is known to have the tendency to be stuck in local extremums, TS address this problem by prohibiting already visited solutions and avoiding the problems of cycles, in this way, the whole solutions space can have the chance to be visited. For more details about the components description and the complete algorithm of TS algorithm integrated in our proposed hybrid approaches please refer to [14],

E. Ant Colony Optimization

ACO algorithm is a probabilistic technique for tackling computational problems which can be reduced to seeking optimal paths in graphs, it was first introduced by Dorigo, Maniezzo and Colorni [15].

The principle of the first algorithm is inspired from the capability of ants to seek the best path from their colony to source food and vice versa. As they move, they deposit an organic compound on the ground called pheromone, paths

(solutions) with a higher pheromone level have a higher probability of being selected by ants (better solutions).

In this paper, the ACO is used in order to diversify the search process, and to explore new regions that may have not been visited in previous iterations by SA algorithm. For more details about the components description and the complete ACO algorithm used in this paper refer to [14].

F. First Hybrid Approach: SA Combined with ACO

In [8], ACO algorithm showed a high diversification ability which allowed it to explore new areas of solutions. To take advantage of this feature, we propose a hybrid approach that combines SA and ACO. The ACO algorithm is used in order to diversify the search process when the SA algorithm can no longer improve the current solution.

In summary: an initial solution will be generated using the Prim method, then the SA will be launched until until that we don't observe any improvement is occurring on the current solution; at that time we call the ACO algorithm to move the search to other regions of solutions space. Next, the SA will resume. Below the outline of this hybrid approach that we note Hybrid SA-ACO :

Step 1. Initialization of SA parameters and generation of an initial solution (see Section III-C).
Step 2. SA and ACO
- a) Launch the SA algorithm as described in Section III-C until $TIME\_TO\_DIVERSIFY$ is reached.
- b) Run the ACO algorithm as described in [8]
- c) Repeat steps 2.a and 2.b until TMP_LEVEL is reached.

$TIME\_TO\_DIVERSIFY$: ACO diversification procedure is launched when this time is reached.

G. Second Hybrid Approach: SA Combined with ACO and TS

In the literature, TS has shown a high intensification potential in finding good solutions in narrow search space for the k-MST problem, the experimentations carried out by [8], have proven this ability, the numerical results showed also that TS is very efficient in case of k-MST with large cardinalities. In [14] TS had had showed another ability which is that it can be a good partner in a hybrid algorithm, the combination with the ACO algorithm had allowed to find new best values for the objective function. In [16] another hybrid approach was proposed by combining TS and SA. These results were obtained even that the neighborhood structures chosen by each approach were different

In this hybrid approach, we combine the three meta-heuristics seen before, namely, SA, TS and ACO, we will exploit the advantages of each of them to find better solutions. This approach starts with an initial solution generated using the Prim method, then the SA will take the hand to improve this solution under the control of the parameter $TIME\_TO\_DIVERSIFY$; the ACO will be launched to take the search to another region of solutions not yet explored. When the stopping criterion (TMP_LEVEL) is reached, the TS is launched on the best solution found to intensify the search in its neighborhood and get the best one.

The outline of this algorithm, that we note Hybrid SA-ACO-TS, is as follows:

Step 1. Initialization of SA parameters and generation of an initial solution (see Section III-C).
Step 2. SA and ACO
- a) Run the SA algorithm until $TIME\_TO\_DIVERSIFY$
- b) Launch the ACO algorithm
- c) Repeat the two previous actions until TMP_LEVEL

Step 3. Tabu Search
- a) Run the TS algorithm as described in [14], the best solution found in Step 2 is the starting solution of this step.

IV. Experimental Results

We have performed numerical experiments using two large regular graphs taking from the benchmark instance KCTLIB. The Table I shows the configuration of these graphs.

TABLE I. Characteristics of the Two Regular Graphs.

Graph name	$\|V\|$	$\|E\|$	Average degree of vertices
Graph 1 : 1000_4_01.gg	1000	2000	4
Graph 2 : g400_4_05.g	1000	2000	4

The results obtained by the proposed approaches are compared to those obtained with the following methods:

- Two solution algorithms proposed in [8], namely TS algorithm and ACO algorithm, denoted by TSB and ACOB, respectively.
- The hybrid solution algorithm proposed in [14], denoted by HybridK.
- The simulated annealing algorithm with restart strategy by [7], denoted by SA.
- The hybrid approach that combine simulated annealing and tabu search by [16].

Our algorithms are coded in C programming language and runned on a computer with a CPU Intel(R) Core(TM) i5, 2.5x2.5 GHz, 4GB RAM. It should be stressed that for TS and ACO algorithms, we have used the same parameter settings as in [14],

Our algorithms have been executed ten times on each graph instance, after that, we record the best, worst, mean and objective function values, also we record the mean time.

Table II presents the SA parameter settings adopted to tackle the graphs 1 and 2. In Tables III- IV results of the proposed approaches are shown. The objective function written in bold style face means that they are best among all obtained values. BNV column gives the best new value of our approaches.

TABLE II. SA PARAMETER SETTINGS ADOPTED TO GRAPHS IN TABLE I.

k	T_0	T_f	TMP_RANGE	TIME_TO_DIVERSIFY	TMP_LEVELS
200	15	0.01	10000	20 s	40
400	15		2000		30
600	15		5000		30
800	10		2000		30
900	10		4000		30

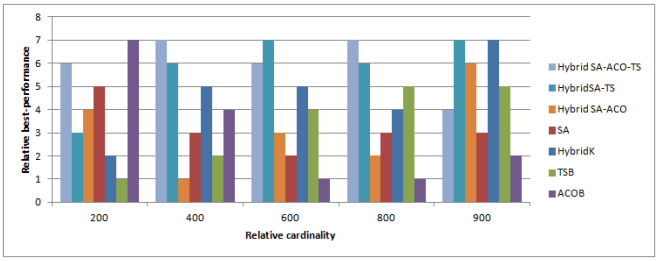


Fig. 3. Performance of Different Approaches for the First Regular Graph.

To easily read and well interpret the Tables III- IV, we transform them into charts, Fig. 3 and 4 represent the order of the performance of each method and for different cardinalities, a higher order value corresponds to better performance.

Fig. 3 show that: for $k = 200$, ACOB is the best; however for other cardinalities we notice that our proposed hybrid approach SA_ACO_TS is very competitive, because we have improved the best known solutions for $k = 400$ and $k = 800$. Fig. 4 show that for $k = 400$, Hybrid SA_ACO_TS is the best; however for other cardinalities we notice that Hybrid SA_TS approach is very efficient in terms of best and mean values. It should be stressed that we have improved the best known solutions for $k = 400$ and $k = 800$. In summary, results reveal that:

- SA is not efficient in case of large graphs.
- SA is very efficient when coupled with TS in finding optimal k-MST solutions.
- TS is known for its great ability to intensify the search, the obtained results had confirmed that; it can be coupled with SA and/or ACO algorithm to obtain good performances.
- ACO is very efficient when its coupled with TS.
- We have achieved good results by hybrid approaches only when TS is part of the hybrid approach, this is due to its high intensification ability.
- Our proposed approaches have consumed more time to provide optimal solutions.

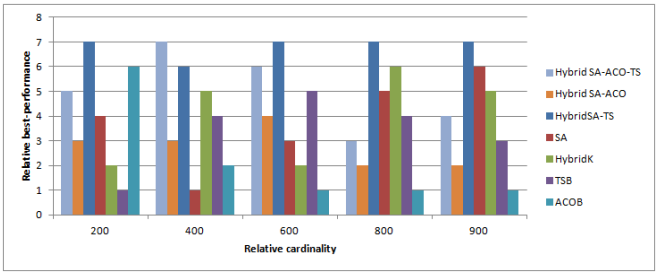


Fig. 4. Performance of Different Approaches for the First Regular Graph.

V. CONCLUSION

This article suggests new approaches to address the k-MST problem. Hybrid approaches combining SA, TS and ACO were presented. In order to show the performance of the these methods, we compared them with other works from the literature using the same benchmark data KCTLIB. The numerical experiments showed that TS is effective to tackle the k-MST problems when it is combined with SA or ACO or both. In our future works, we will focus on how to improve the computational time of our approaches and then address the same problem in case of multi-objective optimization.

REFERENCES

[1] H. W Hamacher, K. Jörnsten, and F Maffioli. Weighted k-cardinality trees, technical report. *Technical Report 91.023*, Politecnico di Milano, Dipartimento di Elettronica, Italy, 1991.

[2] Matteo Fischetti, Horst W. Hamacher, Kurt Jörnsten, and Francesco Maffioli. Weighted k-cardinality trees: Complexity and polyhedral structure. *Networks*, 24(1):11–21, 1994.

[3] Christian Blum and Andrea Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.*, 35(3):268–308, 2003.

[4] Matthias Ehrgott, Horst. W. Hamacher, J. Freitag, and F. Maffioli. Heuristics for the k-cardinality tree and subgraph problems. *Fachbereich Mathematik*, 14(8):24, 1996.

[5] Shun Yan Cheung and A. Kumar. Efficient quorumcast routing algorithms. *Proceedings of INFOCOM '94 Conference on Computer Communications*, pages 840–847 vol.2, 1994.

[6] Freitag J. K Ehrgott M. Tree/k subgraph: a program package for minimal weighted k-cardinality-trees and -subgraphs. *European Journal of Operational Research*, 1(93), 214, 1996.

[7] El Houcine Addou, Abelhafid Serghini, and El Bekkaye Mermri. Simulated annealing algorithm with restart strategy for optimizing k-minimum spanning tree problems. *In EDA 2018*, Business Intelligence & Big Data, RNTI-B-14:321–330, 2018.

[8] Christian Blum and Maria J. Blesa. New metaheuristic approaches for the edge-weighted k-cardinality tree problem. *Computers & Operations Research*, 32(6):1355–1377, 2005.

[9] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science (New York, N.Y.)*, 220(4598):671–680, 1983.

[10] Marc C. Robini and Pierre-Jean Reissman. From simulated annealing to stochastic continuation: a new trend in combinatorial optimization. *Journal of Global Optimization*, 56(1):185–215, 2013.

[11] Linzhong Liu, Haibo Mu, Juhua Yang, Xiaojing Li, and Fang Wu. A simulated annealing for multi-criteria optimization problem: DBMOSA. *Swarm and Evolutionary Computation*, 14:48–65, 2014.

[12] F. Glover. Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, vol. 13, no. 5, pp. 533–549, 1986.

[13] F. Glover. Tabu search part ii. *ORSA Journal on Computing*, vol. 2, no. 1, pp. 4–32, 1990.

[14] Hideki Katagiri, Tomohiro Hayashida, Ichiro Nishizaki, and Qingqiang Guo. A hybrid algorithm based on tabu search and ant colony optimization for k-minimum spanning tree problems. *Expert Systems with Applications*, 39(5):5681–5686, 2012.

[15] M. Dorigo, V. Maniezzo, and A. Colorni. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(1):29–41, 1996.

[16] El Houcine Addou, Abelhafid Serghini, and El Bekkaye Mermri. A hybrid algorithm based on simulated annealing and tabu search for k-minimum spanning tree problems. *In ICOA 2019*, pages 1–6, April 2019.

TABLE III. RESULTS OBTAINED BY EACH APPROACH FOR THE FIRST REGULAR GRAPH.

k	BNV		Hybrid SA_ACO_TS	Hybrid SA_ACO	Hybrid SA_TS	SA	HybridK	TSB	ACOB
200		Best	3336	3374	3375	3372	3393	3438	**3312**
		Mean	3354.5	3420.2	3419.9	3450	3453.1	3461.4	**3344.1**
		Worst	**3373**	3463	3514	3514	3517	3517	3379
		Mean time (s)	742	300	600	300	300	300	300
400	7621	Best	**7621**	7717	7646	7713	7659	7712	7661
		Mean	**7669.6**	7822.4	7689.8	7772.8	7764	7780.2	7703
		Worst	**7739**	7899	7778	7851	7819	7825	7751
		Mean time (s)	6328	300	3468	974	300	300	300
600		Best	12783	12805	**12756**	12858	12785	12801	12989
		Mean	12821.1	12887.4	**12795.4**	12908.1	12836.6	12821.8	13115.6
		Worst	**12864**	12999	12845	12971	13048	12869	13199
		Mean time (s)	21815	1644	11904	1948	300	300	300
800	19065	Best	**19065**	19144	19073	19114	19099	19093	19581
		Mean	19110	19162.3	**19081.6**	19213.7	19101.1	19112.6	19718.7
		Worst	19155	19177	**19095**	19275	19128	19135	19846
		Mean time (s)	7200	2509	3180	1948	300	300	300
900		Best	22845	22942	**22827**	22865	**22827**	22843	23487
		Mean	22851	22962.7	22834.5	23052	**22827**	22859.2	23643
		Worst	22866	22995	22851	23165	**22827**	22886	23739
		Mean time (s)	6827	1440	9263	1029	300	300	300

TABLE IV. RESULTS OBTAINED BY EACH APPROACH FOR THE SECOND REGULAR GRAPH.

k	BNV		Hybrid SA_ACO_TS	Hybrid SA_ACO	Hybrid SA_TS	SA	HybridK	TSB	ACOB
200		Best	3636	3661	**3630**	3639	3667	3692	3632
		Mean	3671	3686.8	**3653**	3699.9	3697.5	3722.0	3670.1
		Worst	3716	3722	**3682**	3784	3738	3751	3710
		Mean time (s)	995	1028	2777	2735	300	300	300
400	8240	Best	**8240**	8359	8292	8378	8323	8358	8376
		Mean	**8293.7**	8393.2	8343.1	8430	8357.1	8385.6	8408.3
		Worst	**8367**	8436	8472	8510	8424	8415	8442
		Mean time (s)	4272	1369.1	4200	1301	300	300	300
600		Best	13681	13743	**13617**	13761	13807	13735	14085
		Mean	13708	13816.4	**13665.6**	13788.2	13824.3	13759.4	14164.5
		Worst	13744	13876	**13690**	13841	13900	13820	14235
		Mean time (s)	28727	2032	16042	2082	300	300	300
800		Best	20149	20208	**20108**	20127	20110	20130	20661
		Mean	20170	20251	20143.6	20169	**20129.9**	20142.9	20811.3
		Worst	20189	20279	20167	20218	**20143**	20155	20940
		Mean time (s)	21355	2739	7980.6	3802	300	300	300
900		Best	24039	24103	**24030**	24032	24035	24044	24782
		Mean	24070	24136	**24034.6**	24045.3	24035	24052.6	24916
		Worst	24090	24172	24040	24052	**24035**	24064	25037
		Mean time	8756	1646	15671.6	4339	300	300	300

[17] F. Glover and M. Laguna. Tabu search. *Handbook of Combinatorial Optimization,Springer,Boston*, pp. 2093–2229, 1998.

Automatic Classification of Preliminary Diabetic Retinopathy Stages using CNN

Omar Khaled[1], Mahmoud ElSahhar[2], Mohamed Alaa El-Dine[3], Youssef Talaat[4], Yomna M. I. Hassan[5], Alaa Hamdy[6]
Faculty of Computer Science,
Misr International University,
Cairo, Egypt

***Abstract*—Diabetes Mellitus is one of the modern world's most prominent and dominant maladies. This condition later on leads to a menacing eye disease called Diabetic Retinopathy (DR). Diabetic Retinopathy is a retinal disease that is caused by high blood sugar levels in the retina, and can naturally progress to irreversible vision loss (blindness). The primary purpose of this imperative research is the early detection and classification of this hazardous condition, to try and prevent any threatening complications in the future. In the course of recent years, Convolutional Neural Networks (CNNs) turned out to be exceptionally famous and fruitful in solving and unraveling image processing and object detection problems for enormous datasets. Throughout this pivotal research, a model was proposed to detect the presence of (DR) and classify it into 5 distinct stages, factoring in an immense and substantial dataset. The model starts by applying preprocessing techniques such as normalization, to maintain the same dimensions for all the images before proceeding to the main processing stage. Furthermore, diverse sampling methods such as "Resize & Crop", "Rotation", and "Flipping" have been tested out, so as to pinpoint the best augmentation technique. Finally, the normalized images were fed into a Convolutional Neural Network (CNN), to predict whether a person suffers from DR or not, and classify the level/stage of the disease. The proposed method was utilized on 88,700 retinal fundus images, which are a parcel of the full (EyePACS) dataset, and finally achieved 81.12%, 89.16%, and 84.16% for sensitivity, specificity, and accuracy, respectively.**

***Keywords*—Diabetes mellitus; diabetic retinopathy; DR; convolutional neural networks (CNNs); image processing**

I. Introduction

A. History and Background

As indicated by the World Health Organization (WHO) [1] around 422 million individuals worldwide have been determined to have Diabetes Mellitus. These cases were especially in low and middle income nations, and the expressed numbers are only expected to increment with time. As indicated by Lee's et al. [2] studies, 33% of individuals experiencing Diabetes Mellitus are likewise determined to have other eye maladies, such as Diabetic Retinopathy. This implies that around 147 million individuals are at risk.

The correlation between Diabetes and retinal complications has been first found and presented in 1856, however, it was not until the second half of the 20^{th} century that this eminent work gave more proof that proposed that Retinopathy really was an entanglement of diabetes.

In recent years a new approach to accurately diagnose and detect the presence of Diabetic Retinopathy has been introduced. The approach mainly replaces the old-fashioned manual diagnosis of (DR), with a modern automated method. Automatic classification and analysis of retinal fundus images is materializing as one of the most significant screening tools for the early detection of (DR). This new approach not only provides more reliable and accurate results, but it also saves a lot of time and money.

Diabetic Retinopathy is an illness that causes retina irregularity from the norm, and in extreme conditions can without a doubt lead to total blindness. A classification technique was suggested that interprets and extracts features and aspects from retinal fundus images, and determines whether an individual experiences (DR) or not, and what level or stage is he/she currently at. This research is additionally centered around distinguishing and immediately perceiving the characteristics and qualities of (DR) for ideal precision during the classification operation.

B. Motivation

Around 39 million individuals in the MENA region (Middle East and North Africa) experience the ill effects of Diabetes Mellitus, and it is without a doubt expected that by 2045 this number will ascend to 67 million. The inspiration and motivation to tirelessly seek after this particular issue was that out of these alarming numbers, 8.2 million cases were in Egypt in 2017 as indicated by the "Worldwide Diabetes Federation"[3]. Which further implies that third of this number is undoubtedly at risk of experiencing the ill effects of (DR).

After thoroughly investigating and breaking down the market, it was found that the preliminary phases or stages of Diabetic Retinopathy and other eye ailments were not identified precisely manually. Moreover, two principle disadvantages that were without a doubt pivotal factors in precisely identifying (DR) were likewise revealed; the datasets utilized in the process were surprisingly little, which obviously prompted the second potential downside that being, low classification and accuracy rates. So based on this critical information, the primary point becomes to devise a successful method for classifying and identifying the preliminary phases of Diabetic Retinopathy, for possible clinical advantages.

1) Problem Analysis: The main issue is that very high blood sugar levels over a broad period of time, cause harm throughout the entire human body. This case occurs when the blood vessels behind the retina get weakened over time and get

damaged, which eventually causes new abnormal blood vessels to grow at the back of the eye. Moreover, these abnormal blood vessels not only leak fluids into the eye, but they can also cause more serious complications such as vision loss (blindness) or glaucoma.

Diabetic Retinopathy is viewed as one of the deadliest infections around the globe, since one probably won't have any visible symptoms of (DR) in the beginning phases, yet as the ailment advances or normally progresses, Diabetic Retinopathy side effects and symptoms might occur. (DR) symptoms and side effects can include Blurred vision, impaired color vision, or spots and dark strings floating in one's vision (floaters).

II. Related Work

The following section introduces the most relevant published work, that primarily depicts and represents the proposed research in terms of applied algorithms, datasets used, number of classified stages, and the overall achieved accuracy.

A. SVM Classifier

Bhattacharjee et al.[4] using Random Forest classifier, classified Diabetic Retinopathy based on three features, which are the area of microaneurysm, the area of blood vessels, and the area of exudates. And using these features, they have classified them into five stages: normal, mild, moderate, severe and Proliferative using Kaggle resized images which are about 10052 images for training, and 3350 images for testing to achieve an accuracy of 76.5%.

Kumar et al. [5] classified 89 images from DIARETDB1 dataset into two stages after removing the noise on the images by using (CALHE) histogram equalization. They also extracted hard exudates, Blood vessels, the area of MA, and the number of MA. Outputting a result for sensitivity and specificity of 96% and 92%, respectively.

Cisneros et al. [6] reached an accuracy between 84.6% and 87.3% by using 413 images for training and 130 for testing, to extract the hard exudates. They also segmented the blood vessels and other properties.

Tjandrasa et al. [7] used soft margin SVM on 149 images from the Messidor dataset, extracting from it the features and properties such as area, perimeter, standard deviation and energy of each exudated image during the feature extraction process, to finally classify between Moderate and Severe cases. They reached an accuracy of 90.54%.

Carrera et al. [8] used 400 images from the Messidor dataset which contains four different stages: Normal, Mild, Moderate and Severe. They then extracted the features of the images by detecting blood vessels, microaneurysms, hard exudates, and other features. They finally reached an accuracy of 85%.

Sangwan [9] trained their system on 96 images while using 54 images to classify three different stages: Mild, Moderate and Proliferative. After performing histogram equalisation on the dataset and turning it into grey-scale images, they reached an overall accuracy of 92.6%.

1) Convolution Neural Networks Algorithms: Lian et al. [10] explored three neural network architectures: AlexNet, ResNet-50 and VGG-16 on a dataset that was provided by EyePACS via kaggle. The dataset consists of 35,126 fundus images and distributed to five classes: normal, mild NPDR, moderate NPDR, severe NPDR and severe PDR. The classes have a Proportion of 73.46% ,6.69 % , 15.06 % , 2.50% and 2.02%, respectively. They then normalized all the images from their original size into 256x256 pixels. Also, they re-sampled the images for the over represented classes, and randomly subsampled the underrepresented classes. Finally, they used spatial translation with one pixel in both left and right horizontal directions, to increase the number of images and avoid bias. The three models achieved accuracy rates of 73.19% for the AlexNet model, 76.41% for the ResNet-50, while VGG-16 achieved the best accuracy which was 79.04%.

Harun et al. [11] classified two stages of Diabetic Retinopathy, using 1,151 fundus images divided into 70:30 data proportion, in which 806 images were used for training, while 345 images were used as testing images. They then achieved an overall accuracy of 67.47% for classifying the two classes, 66.4% and 64.48% for DR and No DR, respectively. They finally used a Multi-layer Perceptron (MLP), trained by Binary relevance for classification with 50 training epochs and 20 hidden layers.

Li et al. [12] proposed a system to classify two stages of Diabetic Retinopathy, and its five different stages using Deep Convolutional Neural Networks. They used a Kaggle dataset divided into 34,124 for training, 1,000 for validation, and 53,572 for testing. They finally reached an accuracy for the five-class classification of 86.17%, while the accuracy for the binary class was 91.05%.

Challa et al. [13] detected and classified the five different stages of Diabetic Retinopathy using an All-CNN architecture that has ten convolutional layers and a Softmax layer. They used a Kaggle dataset divided into 30,000 images for training, and 3,000 images for testing. They then applied some preprocessing techniques on the dataset such as removing black Boundaries, and data augmentation such as vertical and horizontal flipping; rotation in different angles between 45° and 180° to make sure that all the images in stages 1,2,3 and 4 are equal to the images in level 0. They finally achieved an accuracy of 86.64% for classifying the five stages of Diabetic Retinopathy, but from observing the the percentage of Recall, Precision and F1 score in the five classes, class (0) had the highest percentage compared to other classes that didn't exceed 60%.

Junjun et al. [14] applied the Residual Network (ResNet) approach on the EyePACS dataset that contains about 35,126 images, using about 30,000 images for training, and around 5,000 images for testing. The images were then resized to 256 × 256 pixels, and due to the large number of images from class 0, they augmented the images of the other classes to avoid over-fitting, by flipping the images and rotating randomly between 0° and 360° to classify 5 stages with an accuracy of 78.4%.

Jain et al. [15] detected Diabetic Retinopathy and evaluated its severity through using different Convolutional Neural Network (CNN) Architectures such as VGG-16, VGG-19 and

Inception v3 architectures. They divided the 35,126 images from the EyePACS dataset into 60% for training , 20% for validation, and the last 20% to classify 5 stages. The images have been passed on preparation techniques such as Normalization and Data Augmentation, by rotating the images for training by 90° and 270°. As for their results they reached an accuracy of 71.7%, 76.9% and 70.2%, respectively.

Kwasigroch et al.[16] classified 5 stages of Diabetic Retinopathy by using VGG-D architecture on 37,000 images after scaling and cropping the images to 224 x 224 pixels. Data augmentation method was performed on the images, such as horizontal and vertical shifts & flips, rotations, and zooming. They finally obtained an accuracy of 81.7%.

kajan et al.[17] designed a Diabetic Retinopathy classifier to detect the degree of the presence of the disease in the eye using different pre-trained deep neural networks models, such as: VGG-16, ResNet-50, and Inception-v3 on the EyePacs dataset. They used 75% of the dataset randomly for training and the rest for testing, they then created two models, the first model classified the presence of Diabetic Retinopathy in the fundus images, while the second model classified the degree of this disease into four different stages. They achieved their best results of the average classification accuracy of the first classification model using ResNet50 model, and achieved an accuracy of 92.64%. While the other model using InceptionV3 reached an average of 70.29% among the four stages.

Suriyal et al. [18] used MobileNets model for classifying two stages (non-presence & presence) of Diabetic Retinopathy on 16,798 resized images, and used 1000 images for testing from Kaggle dataset. The overall accuracy achieved was 73.3%.

Harangi et al. [19] used different CNN architectures like ResNet, VGGNet, GoogleNet and AlexNet, on about 552 images from e-ophtha-MA, ROC, DIARETDB1, and only 32 images from the dataset Messidor. They also resized the images to 224 x224 pixels to reach an accuracy between 78.56% & 83.35%.

Khan et al.[20] used cropped resized images, and performed histogram equalization on the Messidor dataset before inputting them into these pretrained models: SqueezeNet, AlexNet and VGG-16, to detect the presence of Diabetic Retinopathy. Finally, their classification produced an accuracy between 91.82 % and 94.49 %.

Zeng et al. [21] classified Diabetic Retinopathy using the Kaggle data set as input fundus images, with a training set of only 28,104 images, and a test set of 7024 images. And also weight-sharing layers based on two architecture Inception-V3 pretrained model siamese-like network structure. They also used pre-processing methods like flipping the images horizontally, geometric transforming as cropping, scaling, translating, and shearing the fundus images. Their final result showed that they achieved a score of 82.2%.

Carson et al. [22] classified the disease into four distinct levels using convolutional neural networks (CNN) such as: AlexNet and GoogLeNet models. The dataset used was a mix between the Kaggle dataset of around 35,000 fundus images, and the Messidor dataset of 1,200 fundus images after passing from different preprocessing methods as cropping the images to separate the circular colored image of the retina using Otsu. They also normalized the images using (CLAHE) histogram equalization algorithm, and data augmentation by zooming, rolling and rotating the images to reduce the over-fitting. The final overall accuracy was between 57.2% and 74.5%.

III. Methodology

The main goal of this software is to automatically detect the early stages of Diabetic Retinopathy and classify the level of the disease in the patient's body. Our aim in this project is to help as many Diabetic patients as possible, by preventing Diabetes from affecting their eyesight and progressing to Diabetic Retinopathy. The idea of the system, after thoroughly reading about Diabetes and Eye diseases, was that Diabetic Retinopathy doesn't show any symptoms, until a very late stage in life. As shown in Fig. 1, the system starts by taking a retinal fundus image as input from the diseased patient and apply some data preprocessing techniques. Using a deep learning approach, the system will then detect whether a person suffers from Diabetic Retinopathy or not; based on the answer, the system will then classify the level of the disease and finally propose a solution to the patient.

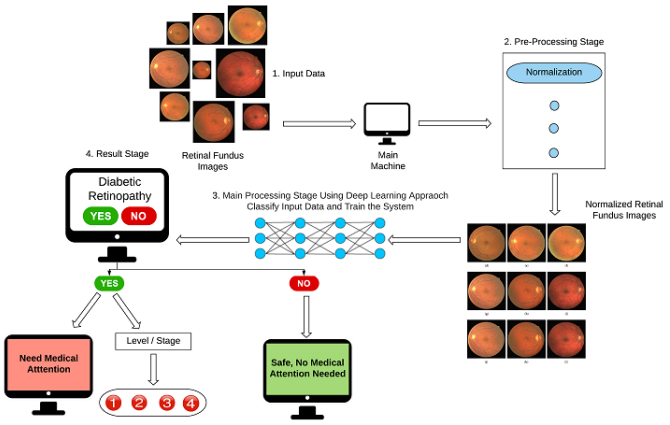


Fig. 1. System Overview.

First, The Main Computer/Machine will collect the information from the Input Images, which are unnormalized Retinal Fundus Images with different sizes chosen from our dataset. The System will then apply some data pre-processing algorithms such as Normalization so that all the images are the same size and dimension. This will thoroughly help us in the main processing phase by simply reducing the complexity of the Input images. Then comes the Main Processing stage in which we test and train our system, We use a Convolutional Neural Network in which a group of connected nodes distributed on multiple layers enhance and strengthen each other along with the Tensor-flow library for feature extraction and classification of the Input images. The System will then proceed to the final stage, which is the Result stage; If the result turned out to be (YES), the System will show the Level/Stage of the disease on a scale of 4-stages, and finally propose that the patient needs medical attention right away. Else (NO) the System will propose that there is no need for medical attention.

The upcoming techniques and operations have been carefully decided upon, to best fit the proposed research, and emphatically improve the overall system performance.

1) Dataset: After thoroughly researching and analyzing the different kinds and sizes of datasets, the following dataset has been specifically chosen, since it contains the largest number of reliable retinal fundus images.

The High-Resolution dataset being utilized in this study is called (EyePACS) [23], which consists of 88,702 images, provided by Kaggle [24], and classified into 5 stages as shown below in Fig. 2. Furthermore, the number of images within each class is shown below in Table I.

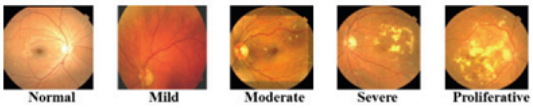


Fig. 2. Stages of Diabetic Retinopathy (DR) with Increasing Severity. [25]

TABLE I. DATASET IMAGES DISTRIBUTION

	Class 0	Class 1	Class 2	Class 3	Class 4
Images	65,343	6,205	13,153	2,087	1,914

A. Dataset Preparation

1) Data Filtration: After examining the dataset, it was found that about 245 images were totally corrupted as shown below in Table II, and could eventually cause distractions and obstructions to the model while training. Accordingly, it was decided to eliminate these images from the model's training process.

TABLE II. DATASET FILTRATION PROCESS

Class Name	Original Images	Original Images (After Filtration)
Class 0 (No DR)	65,343	65,167
Class 1 (Mild)	6,205	6,190
Class 2 (Moderate)	13,153	13,116
Class 3 (Severe)	2,087	2,087
Class 4 (Proliferative)	1,914	1,901
Total	88,702	88,457

2) Data Normalization: After filtering the dataset from all the corrupted images, a new technique called Normalization is then applied. All the Input retinal images maintain different sizes and dimensions, so to overcome this difficulty it was promptly decided to normalize/resize all the retinal fundus images, so they regularly have similar measurements before continuing to the main processing stage. At this point the normalized images keep a unified size of 224 pixels in width and 224 pixels in height, as shown below in Fig. 3.

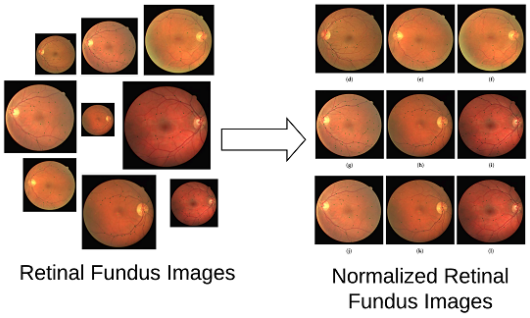


Fig. 3. Retinal Fundus Images Normalization Process.

3) Data Splitting: During the model training process several different splitting techniques were tested out, so as to be able to choose the most suitable, unbiased splitting method.

1) **Random Splitting**: The dataset is randomly split into 70% training and 30% testing, among the 5 different classes shown below in Fig. 4.

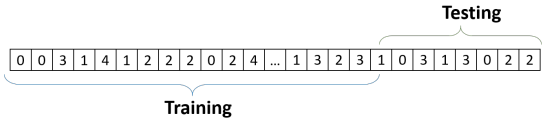


Fig. 4. Random Data Splitting.

2) **Per-Class Splitting**: The dataset is split into 70% training and 30% testing for each individual class. To further ensure that the operation is applied perfectly on each class; the retinal images were grouped together according to their class name, then each group (class) was split into 70% training and 30% testing, shown below in Fig. 5.

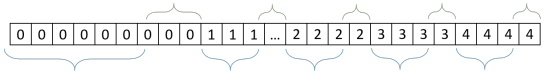


Fig. 5. Per-Class Data Splitting.

3) **Equal Per-Class Splitting**: An equal number of images is taken from each class, and then each class is split into 70% training and 30% among the 5 different classes, shown below in Fig. 6.

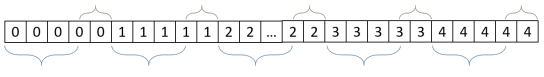


Fig. 6. Equal Per-Class Data Splitting.

4) Conclusion:
After trying out the three different data splitting methods, it was apparent that the "Equal Per-Class Data Splitting" proved to be the best technique, since it ensures that an equal number of training and testing images is used per class, which ultimately guarantees a fair and unbiased system.

4) Data Sampling: As mentioned earlier, the 5 classes within the dataset are unbalanced, and this situation will eventually cause bias towards the largest class, which in this case is class 0 (No DR). To overcome this difficulty, it was decided to apply various up-sampling techniques, so that the number of images per class becomes even. Therefore, to find the most suitable technique, we had to try out three distinct methods ("Rotation", "Flipping", "Resize & Crop").

1) Rotation:
Starting with the Rotation technique, we applied this method through rotating the images by (90, 180, 270) degrees shown below in Fig. 7, so as to be able to generate new image samples.

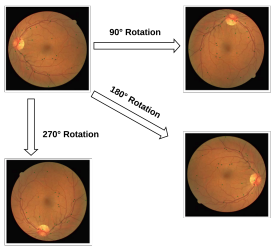


Fig. 7. Retinal Image Rotation.

2) Flipping:
The second method, involves flipping the images horizontally and vertically, shown below in Fig. 8.

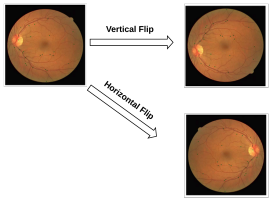


Fig. 8. Retinal Image Flipping.

3) Resize & Crop:
The final technique starts by resizing the images into "384 x 256" images, to maintain the aspect ratio (preventing any loss of data), and then begins cropping random windows of size "224 x 224" to create new dataset samples, shown below in Fig. 9. This technique offered a wide variety of new possibilities than the previous ones, and proved to be the most effective.

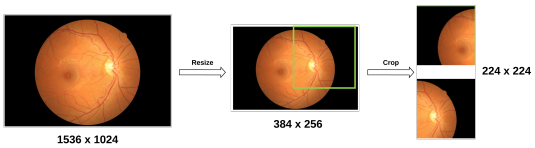


Fig. 9. Retinal Image Resizing and Cropping.

4) Conclusion:
After thoroughly testing out the three different sampling techniques, it was apparent that the "Resize & Crop" method provided the best results, shown below in Fig. 10.

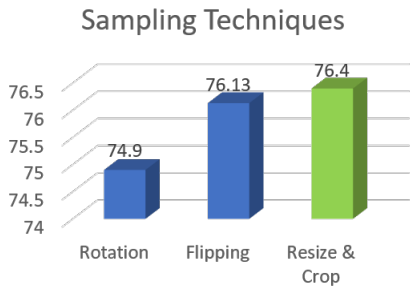


Fig. 10. Sampling Techniques Accuracies.

B. Convolution Neural Network

A Convolutional Neural Network (CNN) is one of the most famous, dominant methods nowadays in object detection and classification, since it doesn't require any feature extraction or segmentation processes, and because it is better at working with enormous datasets. Therefore, a comparison between the distinctive CNN models and methods was conducted, to compare their results and select the most efficient and reliable one out of them.

IV. Experimental Section

The purpose of this experimental section is previewing and discussing the performance and the results of all the trials that have been accomplished throughout this project. To ensure that we get the best output and performance out of our system, we regularly compared and analyzed the results of our trials, so as to be able to figure out the ideal setup and build a more reliable system. Furthermore, the following comparisons were mainly focused on two main aspects; the time taken, and the accuracy achieved. Finally, our conclusion was mainly based on choosing or selecting the parameter that achieved the most reasonable accuracy relative to the time taken.

The following trials represent ten of our most effective trials, with respect to different parameters:

A. Dataset Trials

1) Images Dimensions:

a) Description: This trial aims to illustrate the most suitable images dimensions to use while training the model. This experiment was conducted using three different image dimensions ("224 x 224", "384 x 256", "512 x 512") to conclude the ideal image size that will be used in resizing the training images.

b) Result: As shown in Table III, there was a huge difference between the output of the third trial and the rest of the trials regarding the time taken. Although the "512 x 512" image sizes achieved slightly the highest accuracy, the time taken was excessively large which is not worth it at all. So, "224 x 224" seems to be the best image size to be used due to its low time taken and its reasonable accuracy.

TABLE III. IMAGES DIMENSIONS

Images Dimensions	Time Taken	Accuracy
224 x 224	1:51:47	75.53%
384 x 256	3:27:04	75.13%
512 x 512	13:56:45	76.14%

2) Sampling Techniques:

a) Description: This trial aims to illustrate the most suitable sampling technique to use while training the model. This experiment was conducted using three different sampling techniques ("Rotation", "Flipping", "Resize & Crop") to conclude the most suitable technique to be used in our case. Starting with the "Rotation" technique which depends on rotating the images (90, 180, 270) degrees in generating new image samples. "Flipping" technique performs similarly as well by flipping the images (horizontally, and vertically). Finally, the "Resize & Crop" technique which starts by resizing the images into "384 x 256" images to maintain the aspect ratio (to prevent any loss of data) and then begins cropping random windows of size "224 x 224" to create new samples. This technique offer a wide variety of possibilities than the previous ones.

b) Result: As shown in Table IV, there wasn't a huge difference between the output of these trials. As the "Resize & Crop" technique achieved both the least time taken and the highest accuracy in addition to that it did not cause our model to overfit unlike the other techniques. So, it seems to be the best choice when data sampling is needed.

TABLE IV. SAMPLING TECHNIQUES

Sampling	Time Taken	Accuracy
Rotation	2:48:41	74.97%
Flipping	3:37:13	76.13%
Resize & Crop	2:46:26	76.43%

B. CNN Setup Trials

1) Base Model:

a) Description: This trial aims to illustrate the most suitable base model to use while training the model. This experiment was conducted using four different base models ("VGG16", "VGG19", "Inception V3", "ResNet50").

b) Result: As shown in Table V, there was a huge difference between the output of these trials. In terms of time taken the "Inception V3" model achieved the best timing while its accuracy was considered as the lowest. However, the "VGG16" trial achieved the highest time taken in addition to being the best performing model in terms of accuracy, with a large difference compared to others. So, we believe that the "VGG16" model is the best solution as the output is worth the time taken.

TABLE V. BASE MODELS

Base Model	Time Taken	Accuracy
VGG16	4:22:09	76.40%
VGG19	2:30:44	71.69%
Inception V3	1:50:25	68.00%
ResNet50	2:17:57	62.83%

2) Model Weights:

a) Description: This trial aims to illustrate whether it is better to use "Imagenet" weights while training the model or to train the model from scratch. This experiment was conducted once without any weights and once using "Imagenet" weights to compare between both results.

b) Result: As shown in Table VI, there was a huge difference between the output of these trials, as training from scratch achieved a lower accuracy and took a longer time than using the "Imagenet" weights while training the model, which is obviously the best choice in this case.

TABLE VI. MODEL WEIGHTS

Weights	Time Taken	Accuracy
None	15:06:13	62.50%
Imagenet	4:22:09	76.40%

3) Layers Freezing Techniques:

a) Description: This trial aims to illustrate the most suitable layer freezing technique to be used while training the model. This experiment was conducted using two different techniques ("Without freezing layers", "Freezing layers"). Starting with the "Without freezing layers" approach in which no layers are freezed and the whole model is trained at once, unlike the second approach "Freezing layers" in which the base model layers are freezed at the beginning of the training process and later on unfreezed while training.

b) Result: As shown in Table VII, there was a huge difference between the output of these trials, as the "Freezing layers" approach achieved a higher accuracy and took less time than the "Without freezing layers" approach which is obviously the best choice in this case.

TABLE VII. LAYERS FREEZING TECHNIQUES

Layers Status	Time Taken	Accuracy
No Freezed Layers	12:36:30	62.50%
Freezed Layers	4:22:09	76.40%

4) Batch Size:

a) Description: This trial aims to illustrate the most suitable batch size to use while training the model. This experiment was conducted using three different batch sizes ("16", "32", "64").

b) Result: As shown in Table VIII, there wasn't a huge difference between the output of these different batches. In terms of time taken the "64" batch size achieved the best timing, However, the "32" trial achieved a better accuracy with a minor increase in the time taken than the "64" trial. So, accordingly the "32" batch size was considered the best one.

TABLE VIII. BATCH SIZE

Batch Size	Time Taken	Accuracy
16	4:21:57	75.66%
32	4:22:09	76.40%
64	4:18:48	74.08%

5) Images Distribution within Batch:

a) Description: This trial aims to illustrate the most suitable images distribution within the Batch to be used while training the model. This experiment was conducted using three different distributions ("Consecutive", "Batch", "Block"). Starting with the "Consecutive" approach in which an image from each class is added to the batch in the following order (ex: 012012012...). The "Batch" approach depends on filling the whole batch with images from the same class (ex: 111111111...). Finally, the "Block" approach in which all images from the same class are introduced together before moving on to another class (ex: 0000... - 1111... - 2222...).

b) Result: As shown in Table IX, there wasn't huge difference between the output of these trials. However, the "Consecutive" approach achieved the highest accuracy and best time taken. So, using the "Consecutive" approach seems to be the best way to get the best out of the trained model.

TABLE IX. IMAGES DISTRIBUTION WITHIN THE BATCH

Images Distribution	Time Taken	Accuracy
Consecutive	4:22:09	76.40%
Batch	5:47:09	74.91%
Block	5:42:43	75.46%

6) Optimizers:

a) Description: This trial aims to illustrate the most suitable optimizer to use while training the model. This experiment was conducted using four different optimizers ("adam", "adagrad", "adadelta", "RMSprop") to conclude the most suitable optimizer to be used in our case.

b) Result: As shown in Table X, there wasn't a huge difference between the output of these trials. Although the "adagrad" optimizer achieved slightly the lowest time taken, its accuracy wasn't the best. So, we believe that "adam" is the best choice to be used due to its high accuracy; taking into consideration that its timing was close to "adagrad" timing.

TABLE X. OPTIMIZERS

Optimizers	Time Taken	Accuracy
adam	4:22:09	76.40%
adagrad	4:11:08	75.18%
adadelta	5:33:42	74.98%
RMSprop	4:12:21	75.96%

7) Number of Epochs:

a) Description: This trial aims to illustrate the most suitable number of epochs to use while training the model. This experiment was conducted using four different epochs numbers ("5", "10", "15", "20").

b) Result: As shown in Table XI, there wasn't huge difference between the output of these trials specially for the output accuracy. In terms of time taken the "5" epochs model achieved the best timing, although its accuracy was not the best. However, the "10" epochs trial achieved a slightly higher accuracy, but the time taken was over the double of the "5" epoch trial. So, its obviously clear that "5" epochs is the best solution to be used with respect to these results.

TABLE XI. EPOCHS

No. of Epochs	Time Taken	Accuracy
5	4:22:09	76.40%
10	10:23:08	76.56%
15	12:51:24	75.68%
20	16:20:22	75.55%

C. CNN Architecture

a) Description:: Regarding our CNN design there were two architectures to choose between with respect to their results. First, a "Cascaded Architecture" which consists of two consecutive models "Yes-No Model" which is mainly responsible for detecting the presence of the disease, and "Stages Model" which specify the level of the detected disease by the first model. Second, a one model "5-Stages Architecture" that classifies the presence and the level of disease at once.

b) Result:: Fig. 11 and Fig. 12 represents the confusion matrix of the "Cascaded Architecture" and the "5-Stages Architecture", respectively. To evaluate the performance of the two architectures it was decided to calculate the Accuracy, Sensitivity, and Specificity for each of them.

	0	1	2	3	4
0	4867	24	375	4	52
1	1654	3174	189	197	108
2	2614	843	1167	382	316
3	163	969	305	2722	1163
4	87	318	188	1115	3614

Fig. 11. Cascaded Architecture

	0	1	2	3	4
0	4745	1	518	3	55
1	1624	2905	230	426	137
2	2213	670	1578	544	317
3	119	605	370	3278	950
4	63	157	211	1322	3569

Fig. 12. 5-Stages Architecture

As shown below is Table XII, the "5-Stages Architecture" achieved better results in the three metrics and so this architecture was chosen to be used in our proposed system.

TABLE XII. PERFORMANCE EVALUATION

	Accuracy	Sensitivity	Specificity
Cascaded Architecture	83.37%	58.41%	89.6%
5-Stages Architecture	84.16%	60.41%	90.1%

D. Conclusion

To sum it up, these experiments were applied to monitor the effect of applying various changes to the model's parameters, in addition to analyzing the output results and use them in building a well trained model that will help in providing more reliable results. The below Tables XIII and XIV state the final parameters used and the final results of the system.

TABLE XIII. FINAL MODEL SETUP 1/2

Images Dimensions	Sampling Technique	Base Model	Model Weights	Freezing Technique	Batch Size	Images Distribution
224 x 224	Resize & Crop	VGG16	Imagenet	Freezed Layers	32	Consecutive

TABLE XIV. FINAL MODEL SETUP 2/2

Optimizers	Epochs	Train/Test Percentage	Model Architecture	Time Taken	Accuracy
adam	5	70/30	5-Stages	2 Days, 7 Hrs	84.16%

V. CONCLUSION AND FUTURE WORK

The aim of this proposed system is to be able to automatically detect and classify the various Diabetic Retinopathy stages using a "5-Stages" model architecture, in which deep learning mainly depends on raw colored Retinal Fundus images as its source of input. This system was tested over a total of 26,610 images, which represents almost 30% of the given dataset; after being trained over 62,090 images. As an output, the system achieved an overall accuracy of 84.16% for detecting the presence of the disease and determining its stage. Although, its clear that using these techniques provides a better output compared to the usual machine learning techniques, however, it still requires some extra work to improve these results.

A future work for this paper will be mainly concerned with testing the trained model against real data that has a wide range of variation, to prove its reliability and make sure that this solution is ready to be implemented on real life Diabetic Retinopathy patients. It may also be taken into consideration trying other models, which may offer better results compared to the current ones.

ACKNOWLEDGMENTS

The work done in this research was aided and hugely supported by Dr. Dina Hossam, an assistant professor of ophthalmology at Cairo university. She also offered her assistance with all the needed domain information and knowledge. The authors wish to express their gratitude and appreciation for her huge support.

REFERENCES

[1] Diabetes, (2019, November 25). © 2020 WHO. http://www.who.int/health-topics/diabetes.

[2] Ryan Lee, Tien Y Wong, and Charumathi Sabanayagam. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision*, 2(1):17, 2015.

[3] Diabetes facts and figures, (2019, December 13). © 2020 International Diabetes Federation. https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.

[4] Indronil Bhattacharjee and Tareq Mahmud. *Diabetic Retinopathy Classification from Retinal Images using Machine Learning Approaches*. PhD thesis, 02 2020.

[5] Shailesh Kumar and Basant Kumar. Diabetic retinopathy detection by extracting area and number of microaneurysm from colour fundus image. In *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 359–364. IEEE, 2018.

[6] Fernanda Cisneros-Guzmán, Saúl Tovar-Arriaga, Carlos Pedraza, and Arturo González-Gutierrez. Classification of diabetic retinopathy based on hard exudates patterns, using images processing and svm. In *2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*, pages 1–5. IEEE, 2019.

[7] Handayani Tjandrasa, Ricky Eka Putra, Arya Yudhi Wijaya, and Isye Arieshanti. Classification of non-proliferative diabetic retinopathy based on hard exudates using soft margin svm. In *2013 IEEE International Conference on Control System, Computing and Engineering*, pages 376–380. IEEE, 2013.

[8] Enrique V Carrera, Andrés González, and Ricardo Carrera. Automated detection of diabetic retinopathy using svm. In *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4. IEEE, 2017.

[9] Surbhi Sangwan, Vishal Sharma, and Misha Kakkar. Identification of different stages of diabetic retinopathy. In *2015 International Conference on Computer and Computational Sciences (ICCCS)*, pages 232–237. IEEE, 2015.

[10] Chunyan Lian, Yixiong Liang, Rui Kang, and Yao Xiang. Deep convolutional neural networks for diabetic retinopathy classification. In *Proceedings of the 2nd International Conference on Advances in Image Processing*, pages 68–72, 2018.

[11] Nor Hazlyna Harun, Yuhanis Yusof, Faridah Hassan, and Zunaina Embong. Classification of fundus images for diabetic retinopathy using artificial neural network. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 498–501. IEEE, 2019.

[12] Yung-Hui Li, Nai-Ning Yeh, Shih-Jen Chen, and Yu-Chien Chung. Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network. *Mobile Information Systems*, 2019, 2019.

[13] Uday Kiran Challa, Pavankumar Yellamraju, and Jignesh S Bhatt. A multi-class deep all-cnn for detection of diabetic retinopathy using retinal fundus images. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 191–199. Springer, 2019.

[14] Pan Junjun, Yong Zhifan, Sui Dong, and Qin Hong. Diabetic retinopathy detection based on deep convolutional neural networks for localization of discriminative regions. In *2018 International Conference on Virtual Reality and Visualization (ICVRV)*, pages 46–52. IEEE, 2018.

[15] Anuj Jain, Arnav Jalui, Jahanvi Jasani, Yash Lahoti, and Ruhina Karani. Deep learning for detection and severity classification of diabetic retinopathy. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, pages 1–6. IEEE, 2019.

[16] Arkadiusz Kwasigroch, Bartlomiej Jarzembinski, and Michal Grochowski. Deep cnn based decision support system for detection and assessing the stage of diabetic retinopathy. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 111–116. IEEE, 2018.

[17] Slavomír Kajan, Jozef Goga, Kristián Lacko, and Jarmila Pavlovičová. Detection of diabetic retinopathy using pretrained deep neural networks. In *2020 Cybernetics & Informatics (K&I)*, pages 1–5. IEEE.

[18] Shorav Suriyal, Christopher Druzgalski, and Kumar Gautam. Mobile assisted diabetic retinopathy detection using deep neural network. In *2018 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, pages 1–4. IEEE, 2018.

[19] Balazs Harangi, Janos Toth, and Andras Hajdu. Fusion of deep convolutional neural networks for microaneurysm detection in color fundus images. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3705–3708. IEEE, 2018.

[20] Sharzil Haris Khan, Zeeshan Abbas, SM Danish Rizvi, et al. Classification of diabetic retinopathy images based on customised cnn architecture. In *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pages 244–248. IEEE, 2019.

[21] Xianglong Zeng, Haiquan Chen, Yuan Luo, and Wenbin Ye. Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network. *IEEE Access*, 7:30744–30753, 2019.

[22] Darvin Yi Carson Lam, Margaret Guo, and Tony Lindsey. Automated detection of diabetic retinopathy using deep learning. *AMIA Summits on Translational Science Proceedings*, 2018:147, 2018.

[23] Eyepacs, (2020, January).

[24] Diabetic retinopathy detection, (2019, October 25).

[25] Yuqian Zhou and Shuhao Lu. Discovering abnormal patches and transformations of diabetics retinopathy in big fundus collections. pages 195–206, 01 2017.

Smart Home Energy Management System based on the Internet of Things (IoT)

Emmanuel Ampoma Affum[1], Kwame Agyeman-Prempeh Agyekum[2],
Christian Adumatta Gyampomah[3], Kwadwo Ntiamoah-Sarpong[4], James Dzisi Gadze[5]
Department of Telecommunication Engineering[1,2,3,5],
Kwame Nkrumah University of Science and
Technology, Kumasi, Ghana.

Centre for RFIC and System Technology[4],
School of Information and Communication Engineering,
University of Electronic Science and Technology of China,
Chengdu 611731, People's Republic of China.

***Abstract*—The global increasing demand for energy has brought attention to the need for energy efficiency. Markedly noticeable in developing areas, energy challenges can be attributed to the losses in the distribution and transmission systems, and insufficient demand-side energy management. Demand-oriented systems have been widely proposed as feasible solutions. Smart Home Energy Management Systems have been proposed to include smart Internet of Things (IoT)-capable devices in an ecosystem programmed to achieve energy efficiency. However, these systems apply only to already-smart devices and are not appropriate for the many locales where a majority of appliances are not yet IoT-capable. In this paper, we establish the need to pay attention to non-smart appliances, and propose a solution for incorporating such devices into the energy-efficient IoT space. As a solution, we propose Homergy, a smart IoT-based Home Energy Management Solution that is useful for any market – advanced and developing. Homergy consists of the Homergy Box (which is an IoT device with Internet connectivity, an in-built microcontroller and opto-coupled relays), a NoSQL cloud-based database with streaming capabilities, and a secure cross-platform mobile app (Homergy Mobile App). To validate and illustrate the effectiveness of Homergy, the system was deployed and tested in 3 different consumer scenarios: a low-consuming house, a single-user office and a high-consuming house. The results indicated that Homergy produced weekly energy savings of 0.5 kWh for the low-consuming house, 0.35 kWh for the single-user office, and a 13-kWh improvement over existing smart-devices-only systems in the high-consuming house.**

Keywords*—Internet of things; energy efficiency; home control; smart home

I. Introduction

Energy efficiency has gained as much importance as (if not more than) energy capacity growth. A significant increase in population has resulted in severe electricity supply challenges costing a country like Ghana an average of US $2.1 million in loss of production daily, despite her increase in generation capacity from 1,730 MW in 2006 to 3,795 MW in 2016 [1]. This has resulted in a call for efficient and sustainable energy usage. Efficient energy consumption provides a balance between available energy supply and demand. Energy conservation refers to efforts made to reduce energy consumption which can result in increased financial capital, environmental quality, human comfort, among others [2].

Most processes in place for energy management have been from the side of the supply/distribution, but it is recommended by works like that of [3] that allowing users to manage their electricity usage in an informed manner is a better method. According to [4], a modern view of energy efficiency now requires that low-income economies re-orient themselves toward sustainable energy practices and advanced technology to achieve better energy efficiency. This paper explores the common use of the Internet of Things (IoT) in achieving energy efficiency. IoT technology makes it possible to integrate all devices in the home over networks (including the Internet) for data-sharing (monitoring) and actuation (control). IoT devices are typically integrated with the functionality to enable them communicate over networks and execute certain tasks, and as such are sometimes marketed as being "smart". Non-IoT ("non-smart") devices, despite their proliferation in markets, are not (or have not been made) integrable into the internet-connected IoT space, leaving them out of the general effort to conserve energy.

The novelty in this paper is the integration of both smart and non-smart electrical appliances into the IoT space by designing an IoT device (the Homergy Box) as an intermediary between the smart internet-connected side and all other appliances (including the "non-smart"), such that all devices can be controlled over the internet (in this case, through an internet-connected Android/iOS mobile app).

The architecture of the system will be discussed in Section IV of this paper. Section V describes the communication protocols and hardware/software technologies utilized in this research work, as well as the implementation of the system. The results obtained are described Section VI and conclusions are made in Section VII with avenues for future work.

II. Literature Review

A. Related Works

It is posited by most works that demand-oriented solutions and management systems provide better energy efficiency than

supply-side management systems. The authors in [5] demonstrate that demand-side solutions provide higher efficiency, extra capital cost avoidance, failure probability reduction, risk management improvement, to name a few. Reference [3] bases on such demand-oriented architecture and proposes "a smart domestic energy management system" that connects local appliances to a user interface over the Internet. This was made possible by installing an internet-capable mote on each appliance.

There is general consensus that user inclusion is necessary to achieve better energy management and efficiency [2]. Accordingly, various approaches have been suggested for the implementation of demand-oriented home energy management solutions. The authors in [6] developed and analyzed a 'smart home scheduling scheme' using simulation software, and developed an end-user application interface used for visualizing and controlling energy usage in the home through user-controlled hourly energy-usage schedules. The result showed the avoidance of energy wastage through planning, monitoring and control of daily energy consumption.

In [7]'s intelligent agent-based Home Energy Management System, the authors combine electricity pricing information, smart metering and smart IoT-capable appliances to design a smart system in which users, network operators and energy suppliers can benefit from a dynamic pricing mechanism.

Providing real-time energy usage statistics to users is an important step in improving energy usage efficiency. Various works such as [8] have proposed real-time meter-reading that sends accurate power statistics to both energy suppliers and consumers via intuitive user interfaces. In [9], an IPv6-equipped smart meter prototype reports energy readings to service providers over a cloud. Depending on the user's preference, energy usage in the house adapts to increase/decrease of electricity prices via a gateway which controls appliances. In [8], the authors add actuation and communication functionality to existing traditional meters using an ATMega328P microprocessor (found in Arduino Uno) and a GSM Module. The retrofitted meters in [8] have more accurate meter readings compared to human meter readers and provides consumers with regular updates of their consumption.

Researchers in [10] propose an IoT-based system where a NodeMCU controls relays connected to home appliances. This is a very good solution as it is not specific to smart devices only. However, their work was not deployed in a real environment. Also, they were not exhaustive about the implementation of their proposed voice control. Their solution makes use of a web page interface which may not be as convenient as a native mobile app would be in this case [11].

B. Smart Home Energy Management Systems on the Market

In this subsection we review some of the most common Smart Home Energy Management Systems on the market at the time of writing.

Manufacturers of smart home products usually implement their own control systems and interfaces. These market products usually have a dedicated app for controlling their appliances, such as the ecobee app for controlling ecobee devices. Over time, some manufacturers have opened up their ecosystems to be controlled by other systems ("integration systems"). These systems, such as the Google Home and Amazon Alexa are specifically designed to interface all supported smart devices. The Google Home and Samsung SmartThings are by themselves central control points for an ecosystem of smart devices that support their proprietary integration systems, such as the Google Assistant and Bixby respectively. Integration systems have become common, and manufacturers usually ensure that their smart devices can interface with these systems.

Dedicated hardware devices (known as "smart hubs") have also been developed for integrating smart devices and permitting wider integrations, more convenient user interfacing (such as voice control in smart speakers), and multi-step automation, such as switching lights and Air Conditioners off when the user leaves the house.

III. Motivation and Objectives

From the literature review, it is perceived that a lot of work has been done on Smart Home Energy Systems, but these works have been focused on already-smart appliances. The solutions reviewed do not consider devices that are not IoT-capable (that is, devices that do not have the ability to sense and communicate with their environment). Smart devices, according to [12], "are autonomous physical/digital objects augmented with sensing, processing and network capabilities. They carry chunks of application logic that let them make sense of their local situation and interact with human users. They sense, log and interpret what is occurring within themselves and the world, act on their own, intercommunicate with each other and exchange information with people".

According to Statista's 2020 estimate in Fig. 1, only 0.6% of households in Africa own at least one smart device, compared to 69% in the United States. It can be concluded that markets with a low penetration of smart devices (such as Africa) would not benefit from the proposed solutions reviewed in Section II. The proposed architecture for domestic appliances in [3] (including the non-smart appliances) has a major limitation: an IoT mote will have to be installed for every non-smart appliance. This will come as very expensive and hence impractical for low-income households who arguably need energy efficiency the most. The authors in [2] report a mean electricity usage efficiency of 63% in Ghana, suggesting the existence of an immense potential for the implementation of energy efficiency measures. We are motivated to design and implement a smart energy-efficient IoT-based Home Energy Management Solution relevant to homes in developing areas. Our objectives therefore become clear here as follows:

- To design a smart Home Energy Management System that includes non-smart appliances into the energy-efficient IoT space.
- To implement the proposed system under realistic test environments for possible use-case scenarios.
- To test the effectiveness of the proposed IoT-based solution at improving energy efficiency.

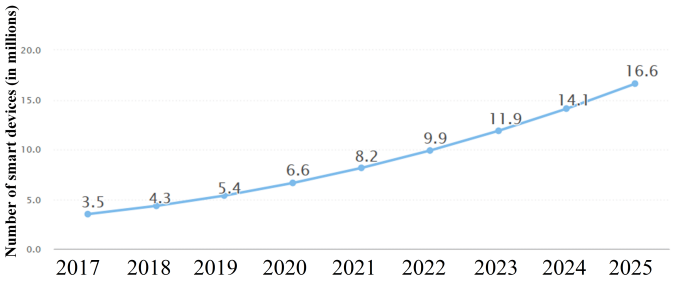


Fig. 1. Penetration of Smart Devices in the African Market. (Statista, 2020)

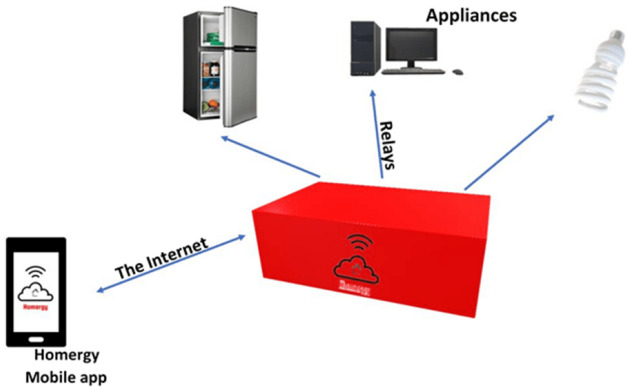


Fig. 2. System Model.

IV. Proposed System Architecture

A. System Model

To make the product relevant to homes with non-smart devices and appliances, the "Homergy Box" was developed to be installed in homes/offices and serve as a gateway between the "smart side" (the Internet and Homergy application), and the "non-smart side" (the home appliances), therefore bridging the gap between these classes of appliances. The model is shown in Fig. 2.

The components in the Box consist mainly of an Arduino Mega microcontroller, a NodeMCU, relay modules and an AC-DC converter. Since the electronics require a constant DC voltage, the AC-DC converter converts the input 240 V AC to 5V DC. This 5V DC is then to supplied to microcontroller, NodeMCU and the relay modules. The NodeMCU has an inbuilt Wi-Fi hardware and connects the Homergy Box to the Internet through Wi-Fi.

The NodeMCU receives data from the mobile app, parses the data and sends appropriate instructions to the Arduino Mega. The Arduino Mega microcontroller executes instructions for controlling the appliance-connected relays. Although the NodeMCU alone could have been used as microcontroller, it does not have enough I/O pins and output power to control the Homergy Box's sixteen (16) relay pins. The relay modules serve to interface the high-voltage AC appliances and the low-voltage electronics circuitry. Fig. 3 shows a more-detailed block diagram of the system.

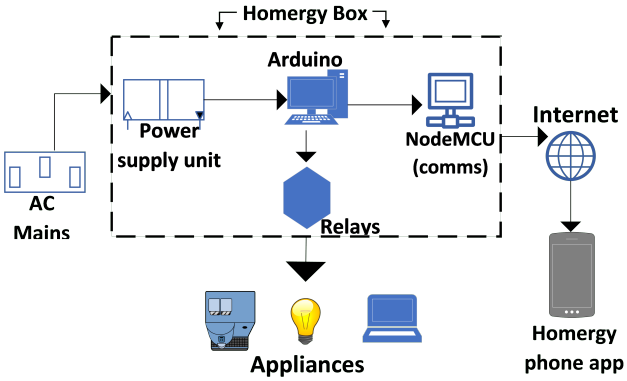


Fig. 3. General System Architecture.

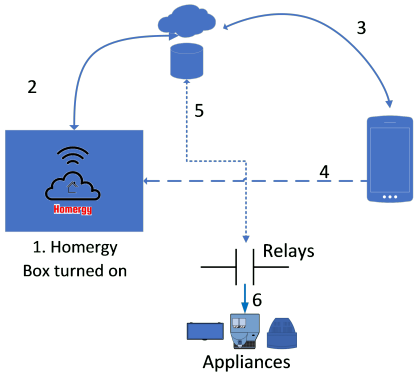
B. Communication

Communication protocols used in Homergy include the I^2C serial protocol, HTTP, WebSocket and TCP/IP. I^2C is implemented between the microcontroller (Arduino) and the Wi-Fi module (NodeMCU), and also between the microcontroller and Liquid Crystal Display (LCD). The user interface (Homergy Mobile App) sends the user's instructions to the cloud-based database via an HTTP request. The database then sends a change event notification to the NodeMCU through an already-established WebSocket over HTTP. The NodeMCU parses the received data and sends the instruction to the Arduino Mega. On reception, the Arduino Mega turns the specified relay on or off.

1) I^2C Communication Protocol: I^2C (pronounced i-squared-see or i-two-see) is a serial communication protocol usually implemented in microcontrollers, EEPROMs, analog-digital converters and sensors. I^2C is a multi-master and multi-slave protocol which makes use of only two wires: a data line called Serial Data (SDA), and a Serial Clock (SCL) line.
For each communication session, the respective master generates a start condition, supplies clock for the communication on the SCL line and also specifies which slave can talk on the SDA line by first sending the slave's address. Therefore, each master or slave must have a unique address. At the end of a communication session, the master generates a stop condition. Data transfer between master and slave is split into 8-bit packets.

2) HTTP: In standard Hypertext Transfer Protocol (HTTP), a client sends a request to a server for data. The server responds to the client by sending the requested data or any error message and closes the connection. Communication only begins when the client first opens a connection and requests data from the server.

3) WebSocket Protocol: A WebSocket is a full-duplex communication protocol which runs on a TCP connection. The connection is not closed after the server sends a response to the client. This allows the server and client to communicate at any time until either of the two closes the connection.



1. Homergy is turned on by user.
2. Homergy reads the Database and initially applies existing values to appliances through relays. A real-time streaming connection is also established with database.
3. The Homergy app, while running, establishes a stream connection with Realtime Database.
4. A virtual stream is established between the Homergy Mobile App and the Homergy Box through Firebase Realtime Database. Phone commands are sent to database.
5. The Realtime Database immediately forwards relay changes (initiated in the app) to the Homergy Box's relays.
6. The relays control the connected appliances.

Fig. 4. Flow of Events.

V. Implementation

This section discusses the implementation of our proposed architecture in Section IV. The user system's operation and user interaction is described in Fig. 4. The following sub-sections are a detailed description of the implementation of specific parts of our proposed architecture.

A. Tools Used

Arduino IDE: Atmel-based boards were used as micro-controllers. The Arduino IDE was used to write and upload programs to the Arduino Mega 2560 board. The Arduino IDE was also used to program the NodeMCU using the "ESP8266 Core for the Arduino IDE" open-source library.

Android Studio: Android Studio was the IDE used to develop the mobile application for this project using the Dart programming language and the Flutter framework.

Cloud platform: A NoSQL Database with real-time streaming capabilities was used to store data on users and Homergy Boxes. In our implementation, the Firebase Realtime Database was selected due to well-documented libraries available for both Android/iOS/Web platforms and the NodeMCU platform. The streaming feature of the Firebase Realtime Database was used to send commands to Homergy Boxes in real time. Any cloud-based implementation (e.g. MQTT used in [13]) that allows real-time communication between the NodeMCU and the Homergy Mobile App will also work.

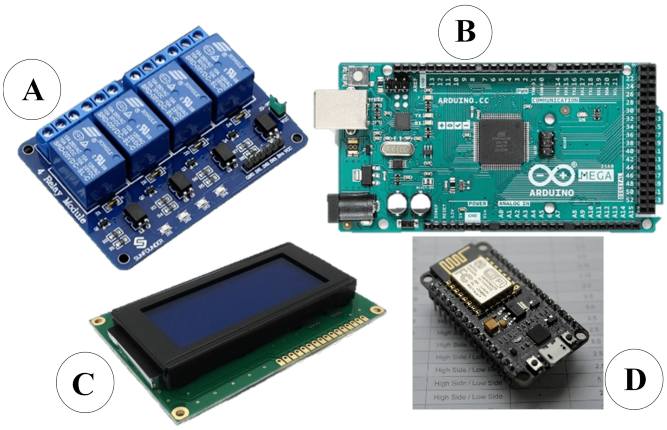


Fig. 5. Hardware Components of the Homergy Box.

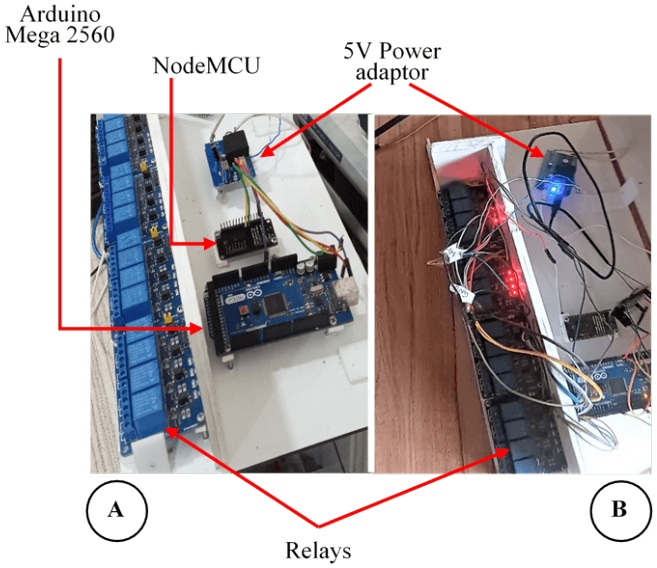


Fig. 6. Internals of Homergy Box with (B) and without (A) Relay Wiring.

B. Hardware Design

The exterior of the Homergy Box hardware is shown in Fig. 15. The internal components are shown in Figs. 5 and 6. Below are the properties of the hardware used in the Homergy Box;

1) Four-channel Relay board (Fig. 5-A): There are four of these relay modules in the Homergy Box, totalling sixteen (16) relay channels. The relay module serves as coupling between the high-voltage home circuit and the low-voltage Homergy Box circuit. The module has its internal low-voltage digital signal circuit isolated from the relay through opto-coupling. The relay's contact capacity is 10A 250V AC / 10A 30V DC, whilst the Digital circuit operates at 5V 20mA (DC). The module is therefore compatible with the Arduino's 40mA General-Purpose Input/Output (GPIO) pins, and with common household AC appliances.

The relays are connected to the appliances as shown in Fig. 7. The Normally Closed (NC) port of the relay is placed between the source and the load. This way, current can still

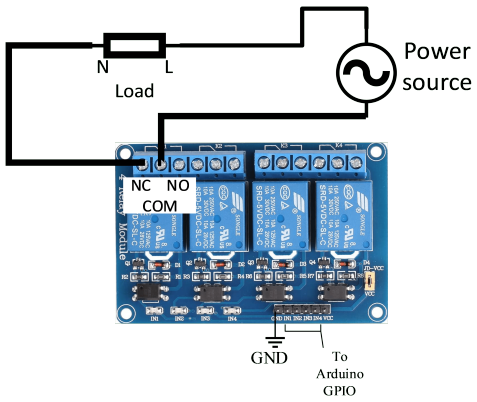


Fig. 7. Normally-closed Relay Connection.

flow when the Homergy Box is off. A "High" from the Arduino GPIO pin activates the relay and switches the appliance off.

2) Arduino Mega 2560 (Fig. 5-B): The Arduino's properties are detailed in Table I.

TABLE I. PROPERTIES OF THE ARDUINO MEGA MICROCONTROLLER

Property	Value
Operating voltage:	5 V
Input voltage:	7 - 12 V (20 V max)
Digital I/O Pins:	54
Memory:	8 kB (Flash), 4kB (EEPROM)
Clock speed:	16 MHz

3) 16x04 I^2C LCD (Fig. 5-C): An LCD is included to provide information on the state of the Homergy Box to the user. During initial configuration, the LCD guides the user in step-by-step procedure to configure the Wi-Fi connection (SSID and password). Any errors (such as disconnection from the Wi-Fi or the Cloud) is displayed on the LCD. The LCD is controlled by the NodeMCU.

4) NodeMCU (Fig. 5-D): Table II shows the properties of the NodeMCU module used.

TABLE II. PROPERTIES OF THE NODEMCU WI-FI MODULE

Property	Value
Operating voltage:	3.3 V
Input voltage:	7 - 12 V
Silicon-on-a-Chip (SoC):	ESP8266 (LX106)
GPIO Pins:	17
Memory:	64 kB (SRAM), 4MB (Flash)
Clock speed:	16 MHz
Wi-Fi band:	2.4GHz only
Operating Modes:	(Simultaneous) Access point (AP) and Station (STA) mode

C. Software Design

There are three main parts of the Homergy system; The Microcontroller (Arduino), the Communications (NodeMCU/ESP8266 and Cloud) and the Homergy Mobile App.

Arduino: The Arduino makes use of the SoftwareSerial library to communicate with the NodeMCU over a serial connection. The Arduino has been programmed to always listen to the NodeMCU for instructions. These instructions are received as a JavaScript Object Notation (JSON) objects which are parsed and executed. The JSON was designed to have two fields: a "command" field and a "payload" field. Command contains a bool which indicates whether the NodeMCU is to change or read the Arduino-connected relays' states. The payload field contains data on user-defined relays states from the Homergy Mobile App, and this field is null only when the NodeMCU is requesting relay states. The Arduino reads or writes the relay states according to the JSON sent by the NodeMCU.

NodeMCU/ESP8266: As shown in Fig. 4, the NodeMCU (bearing an ESP8266 Wi-Fi module) was programmed to connect to the NoSQL cloud-based database (the Firebase Realtime Database) upon start-up. Commands for each Box are saved in a separate child node called "relays" for each Homergy Box (as seen in Fig. 11). The relay states are boolean values (True or False). The Homergy Mobile App, when a user makes a change (on or off), sets the boolean value to True or False respectively. The NodeMCU instantly receives the data for the changed child node (due to Firebase streaming) and sends an appropriate command to the Arduino, which in turn actuates the relay.

D. Homergy Mobile App

The Homergy Mobile App was developed using the cross-platform app development framework Flutter, making it available for major mobile platforms (Web, iOS and Android). The Mobile App was designed to have a nice user-friendly interface. Using the Firebase Database, a rewards system was developed where users gain points (called H-points) for using the system (Fig. 8). Such a reward system is proven to contribute to user retention and engagement [14], and hence leads to better energy efficiency for Homergy users. The app screen on which appliances (relays) are controlled is shown in Fig. 9. The default relay names (Relay 1, Relay 2, etc.) can be edited by the user to more recognizable names (e.g. Projector, Microwave), depending on how the relays are connected to the building by the electrician.

Cloud: The Google Firebase Realtime Database was used as the cloud provider in our implementation. Google's Firebase Realtime Database is a NoSQL JSON-based database with real-time data streaming capabilities. The JSON was structured as shown in Fig. 11. By design, the Firebase sends data change notifications to all clients that have subscribed to listen to a particular node. This feature was taken advantage of. Clients (the Homergy Box and the Homergy Mobile App) subscribe to nodes corresponding to the state of the relays of the user's Homergy Box, and hence receive a real-time payload of the node anytime the user sends a command. This way, the Homergy Box and the Homergy Mobile App virtually communicate directly.

E. Security

Database access rules (Firebase Rules) and a security-focused database structure (Fig. 11) was applied such that each Homergy Box could only be controlled by their authorized users. Email-password authentication is required by the app for

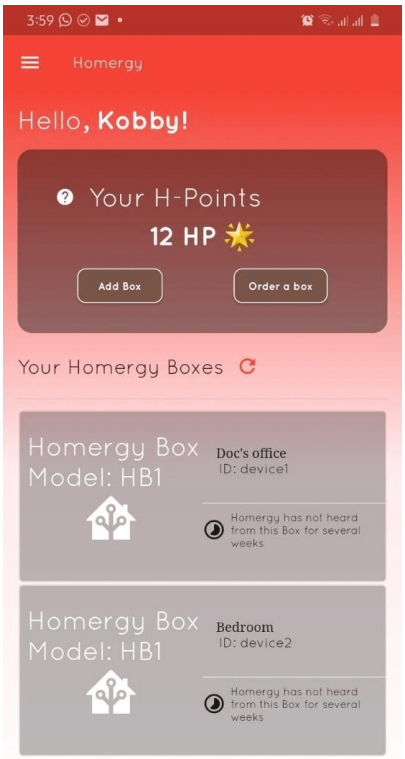


Fig. 8. Home Screen.

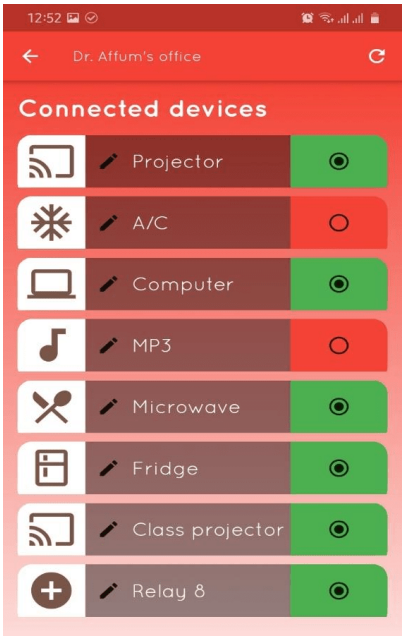


Fig. 9. Box Configuration Page.

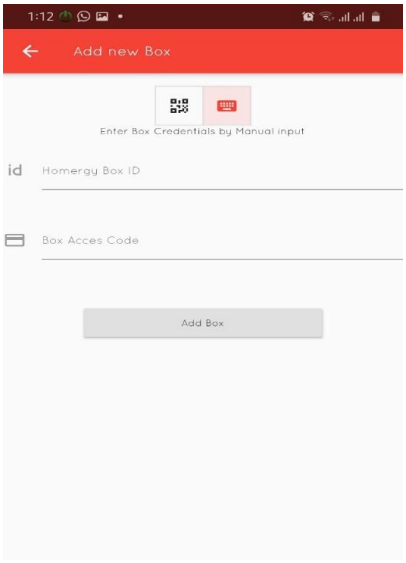


Fig. 10. "Add Box" Page.

Fig. 11. JSON Structure of Firebase Realtime Database.

usage, and users must always re-enter these credentials after an app reset, re-install, or user-reported suspicion of malicious activity. Each Homergy Box has a unique identification, called the "Homergy Box ID" which can be given to each user upon purchase (to give the user access to the Box). The Homergy Box ID maps to a unique Access Code (like a password) which can be obtained by scanning a QR code hidden at the back of each Homergy Box. The QR Code is an encrypted version of the Homergy Box's unique Access Code. After an in-app QR Code scan (Fig. 10), the code is decrypted by the app

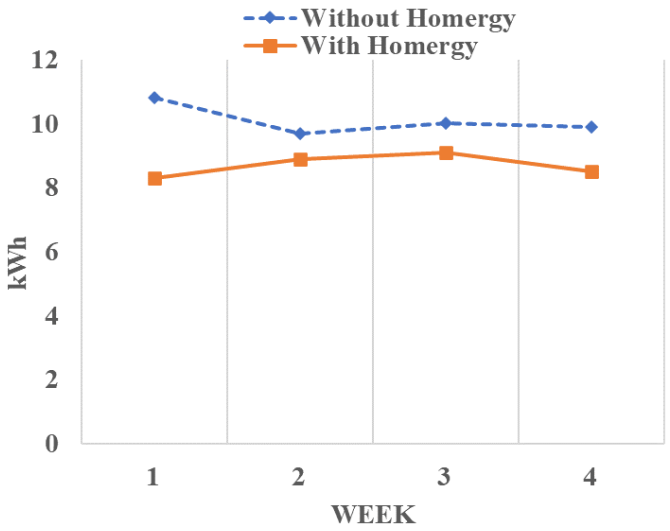


Fig. 12. Homergy vs. no HEMS for a Lifeline Consumer.

and verification and authorization is done. The user can then control the Homergy Box. A user who has been authorized to access a Homergy Box can change the Homergy Access Code with an email-verification procedure, after which the QR Code becomes void. This will also require all other users to sign in again (with the new credentials) to maintain control of the Homergy Box.

Other users in a home can get control simply by scanning the scanning QR Code, or by manually entering the Homergy Box ID and Access Code, within the "Add Box" section of the app, as shown in Fig. 10.

VI. Results and Discussions

In this section, we discuss the results of implementing the proposed Homergy system. The Results of using Homergy as a Home Energy Management System (HEMS) is explored and presented in Fig. 12–14.

Three varied environments were used to measure the effectiveness of Homergy. Two of these environments were selected according to Ghana's energy provider's consumer types: Lifeline Consumers and Non-lifeline Consumers. According to the Public Utilities and Regulatory Commission of Ghana, a lifeline consumer is any consumer whose monthly electricity consumption is less than 50kWh. Consumers with monthly consumption above 50kWh are Non-lifeline consumers. The third environment was a single-user office with Air Conditioning. This is a common energy usage scenario found in workplaces.

Our Key Performance Indicator is the power consumption (in kWh) per week. The kWh parameter has been successfully used by researchers in [2] to compare the energy efficiency levels of various consumer groups. For each of the three environments, we measured the kWh/week for eight weeks. In the first four weeks, there was no Homergy in the environment. In the subsequent four weeks, Homergy had been installed in the environment. Users were allowed to get accostumed to the Homergy Box for a week before measurements were taken in the four subsequent weeks of Homergy usage. The

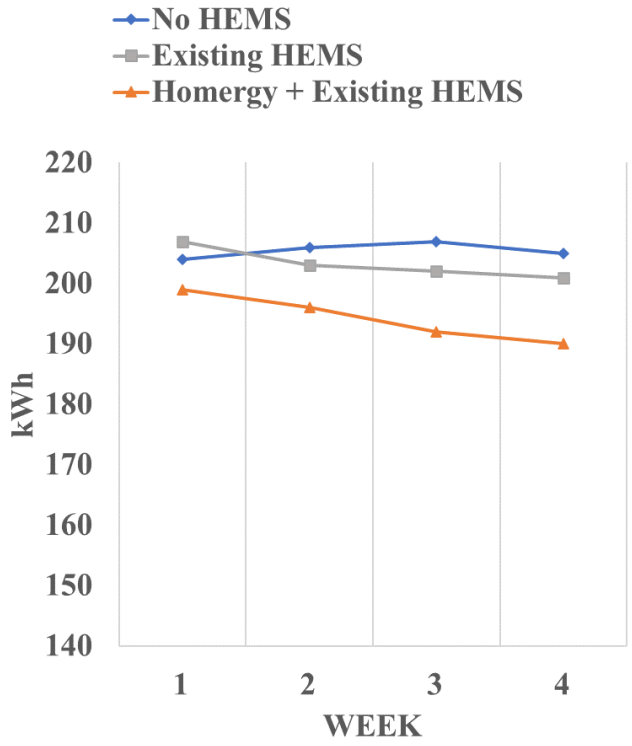


Fig. 13. Homergy vs. Existing HEMS vs. No HEMS for a Non-lifeline Consumer.

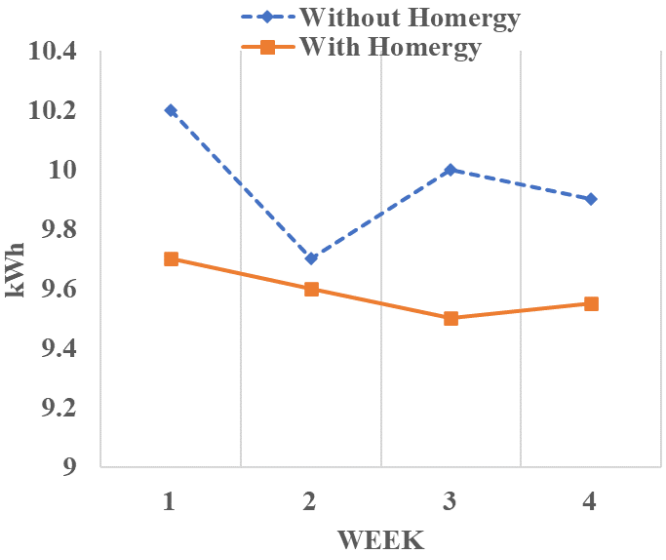


Fig. 14. Homergy vs. no HEMS for a Single-user Office with A/C.

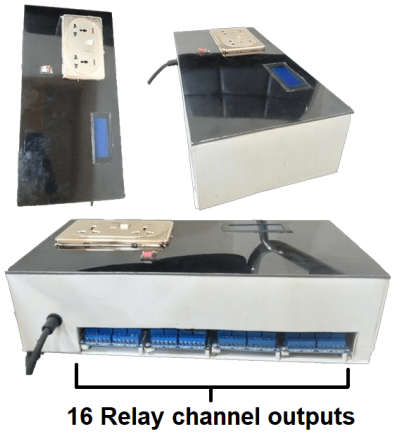


Fig. 15. Homergy Box Hardware.

Non-lifeline consumer had an existing Smart Home Energy Management System (HEMS) whose impact was also measured. The existing HEMS provided energy savings only through smart devices in the house, including smartphone-controlled Air Conditioning, smart motion-activated bulbs, and a smart energy-saving refrigerator. The findings are presented in Fig. 12 – 14. The most commonly used appliances for each environment are shown in Table III.

All three environments with Homergy showed a weekly reduction in kWh compared to their consumption without Homergy. Since the number of users and appliances was constant during the entire test period, this can be explained as an increase in energy efficiency [2].

The implementation of Homergy translated to weekly energy savings of 0.5 kWh for the lifeline consumer, 0.35 kWh for the single-user office, and 18 kWh for the non-lifeline consumer (compared to their usage without any HEMS). According to Ghana's electricity prices at the time of writing, these numbers are equivalent to yearly monetary savings of at least USD 1.3 for the lifeline consumer, and USD 121 for the non-lifeline consumer.

TABLE III. MOST-USED APPLIANCES IN TEST ENVIRONMENTS

Appliance	Power (W)	Quantity		
		Lifeline	Single-user office	Non-lifeline
Water Heating	5500	0	0	4
Electric Kettle	2200	0	1	2
Electric stove	1500	0	0	1
A/C	1100	0	1	4
Washing machine	1000	0	0	1
Fridge	1000	1	1	3
Electric iron	1000	0	0	1
Coffee Maker	800	0	1	1
TV/Large Monitor	75 - 300	1	1	4
Computer/console	80 - 100	0	2	3
Fan	75 - 100	1	1	8
Lighting (LED)	10	3	1	16
Phone charger	5	1	1	4

VII. CONCLUSION AND FUTURE WORK

In this paper, a modular IoT-based Home Energy Management System was successfully developed, giving users the ability to control the electrical consumption of their appliances. This finished work provides users the convenience of controlling appliances from anywhere, integrate the proliferated non-smart devices into the energy-efficient IoT space and also provide a modern energy management approach for both urban and rural areas. The proposed system was also successfully deployed in three real environments that are reflective of possible use cases (moderate energy consumer, offices, and high energy consumers). The results of Homergy's deployment showed increase in energy efficiency for all scenarios (including a 25-kWh energy savings in the first month for a high consumer).

In the future, the Arduino Microcontroller could be removed altogether to make the module cheaper and simpler. This is because the NodeMCU is a microcontroller and can handle the basic switching functions given to the Arduino in our model. The need for more GPIO pins for more relay connections can be met by using special pin-extension modules or shift registers.

The system could be made smarter by taking advantage of Machine Learning. With Machine Learning, the Homergy system could "learn" the habits of users connected to a Homergy Box, and automatically control appliances or give reminders in cases where the users may have forgotten to do so.

ACKNOWLEDGMENT

The authors would like to thank the IDRC's "Strengthening Engineering Ecosystem in Sub-Saharan Africa (SEESA) Project" for financial support provided under IDRC Grant Number: 108883-003, and the technical assistance in carrying out this work successfully.

REFERENCES

[1] Ebenezer Nyarko Kumi, *The Electricity Situation in Ghana: Challenges and Opportunities*. CGD Policy Paper. Washington, DC: Center for Global Development.

[2] D. K. Twerefou and J. O. Abeney, "Efficiency of household electricity consumption in Ghana," Energy Policy, vol. 144, p. 111661, Sep. 2020.

[3] Dlodlo Nomusa, Smith Andrew, Montsi Litsietsi and Kruger Carel, *Towards a demand-side smart domestic electrical energy management system*, 2013. IST-Africa Conference and Exhibition, IST-Africa 2013. 1-12.

[4] Angeliki N. Menegaki, Stella Tsani, *Critical Issues to Be Answered in the Energy-Growth Nexus (EGN) Research Field, The Economics and Econometrics of the Energy-Growth Nexus*, Chapter 5. Academic Press, 2018, pp. 141-184.

[5] Jabir, Hussein & Teh, Jiashen & Ishak, Dahaman & Abunima, Hamza. (2018). *Impacts of Demand-Side Management on Electrical Power Systems: A Review*.

[6] Alimi, O. & Ouahada, Khmaies. (2018). *Smart Home Appliances Scheduling to Manage Energy Usage.* pp. 1-5.

[7] B. Asare-Bediako, W. L. Kling and P. F. Ribeiro, *"Integrated agent-based home energy management system for smart grids applications,"* IEEE PES ISGT Europe 2013, Lyngby, 2013, pp. 1-5.

[8] A. S. Metering, S. Visalatchi and K. K. Sandeep, *"Smart energy metering and power theft control using arduino & GSM,"* 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017, pp. 858-961.

[9] J. Höglund, J. Eriksson, N. Finne, R. Sauter and S. Karnouskos, "Event-driven IPv6 communication for the smart grid infrastructure," *2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, Barcelona, 2011, pp. 1-2.

[10] S. K. Vishwakarma, P. Upadhyaya, B. Kumari and A. K. Mishra, *"Smart Energy Efficient Home Automation System Using IoT"*, 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India, 2019, pp. 1-4.

[11] Jobe, William. (2013). Native Apps Vs. Mobile Web Apps. International Journal of Interactive Mobile Technologies (IJIM). 7. pp. 27-32.

[12] G. Kortuem, F. Kawsar, V. Sundramoorthy and D. Fitton, *"Smart objects as building blocks for the Internet of things,"* in IEEE Internet Computing, Vol. 14, No. 1, pp. 44-51.

[13] Y. Upadhyay, A. Borole and D. Dileepan, *"MQTT based secured home automation system"*, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-4.

[14] Claussen, Jörg & Kretschmer, Tobias & Mayrhofer, Philip. (2012). The Effects of Rewarding User Engagement – The Case of Facebook Apps. Information Systems Research.

Security, Privacy and Trust in IoMT Enabled Smart Healthcare System: A Systematic Review of Current and Future Trends

Thavavel Vaiyapuri[1], Adel Binbusayyis[2*], Vijayakumar Varadarajan[3]

College of Computer Engineering and Science[1,2],
Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia
Centro de Tecnologia,Federal University of Piauí, Brazil[3].

***Abstract*—In the past decades, healthcare has witnessed a swift transformation from traditional specialist/hospital centric approach to a patient-centric approach especially in the smart healthcare system (SHS). This rapid transformation is fueled on account of the advancements in numerous technologies. Amongst these technologies, the Internet of medicals things (IoMT) play an imperative function in the development of SHS with regard to productivity of electronic devices in addition to reliability, accuracy. Recently, several researchers have shown interest to leverage the benefits of IoMT for the development of SHS by interconnecting with the existing healthcare services and available medical resources. Though the integration of IoMT within medical resources enable to revolutionize the patient healthcare service from reactive to proactive care system, the security of IoMT is still in its infancy. As IoMT are mainly employed to capture extremely sensitive individual health data, the security and privacy of IoMT is of paramount importance and very crucial in safeguarding the patient life which could otherwise adversely affect the patient health state and in worse case may also lead to loss of life. Motivated by this crucial requirement, several researchers in tandem to the advancement in IoMT technologies have continuously made noteworthy progress to tackle the security and privacy issues in IoMT. Yet, many possible potential directions exist for future investigation. This necessitates for a complete overview of existing security and privacy solutions in the field of IoMT. Therefore, this paper aims to canvass the literature on the most promising state-of-the-art solutions for securing IoMT in SHS especially in the light of security, privacy protection, authentication and authorization and the use of blockchain for secure data sharing. Finally, highlights the review outcome briefing not only the benefits and limitation of existing security and privacy solutions but also summarizing the opportunities and possible potential future directions that can drive the researchers of next decade to improve and shape their research committed on safe integration IoMT in SHS.**

***Keywords*—Smart healthcare system; internet of medical things; authentication and authorization; security and privacy; blockchain; intrusion detection system**

I. Introduction

In recent years, SHS have greatly increased the economy and is considered as essential component of economy. The IoMT performs a crucial role towards the development of SHS by enabling to develop wide range of applications, say telemedicine, smart medication, onsite and remote monitoring of medical resources. patients treatment compliance and behavioral change. The medical devices which are equipped with sensors and interconnected in the healthcare sector are named as IoMT [1]. The workload in hospitals could be decreased by restricting unnecessary hospital visits with the use of IoMT. Also, it provides a safe data transmission environment for interchanging sensitive medical data amongst diverse medical sectors. IoMT's applications have made lives appropriate. The concern of IoMT security, privacy and trust occurs rapidly. security, privacy and trust have recently received more attention among researcher community [2].

In data security, the storage and transmission of the data is secured and safeguarded to ensure the integrity, validity and importantly authenticity of the data. Further, it assures that the data can be viewed and modified only by the authorized users. Privacy-preserving (PP) is another key objective to be considered while designing an SHS. It mainly account for severity and sensitivity of shared data when it is transmitted over an open and insecure channel. PP involves content and contextual requirements. The patient information is protected against any data leakage by content privacy but achieving patient privacy is a challenge because an attacker can recognize patient health state based on the attended doctor's identity. Also, it is crucial to ensure contextual privacy. Contextual privacy involves of protecting the communication's context. In IoMT enabled SHS, various symmetric and asymmetric encryption method are used to achieve privacy [3].

Recently, it is reported in literature that it is not an optimal solution to apply complex machines learning (ML) algorithms on resource-constrained devices such as IoMT [1]. Yet, it can be resolved by deploying simple PP methods on IoMT devices and utilizing the benefits of cloud for complex ML algorithms [4], [5]. Many works are reported in literature based on cloud related securing solutions for IoMT in SHS. This paper aims to introduce the IoMT enabled SHS architecture, summaries various security, privacy and trust mechanisms published in recent years for IoMT enabled SHS. Finally, it concludes presenting few recommendations for future research directions.

II. Review on Security Mechanism for IoMT Enabled SHS

Security of IoMT in SHS plays a significant role when compared to typical IoT-based infrastructures. Recently, extensive research had been done for securing IoMT enabled smart

TABLE I. COMPARISON OF RECENT PROMISING SECURITY MECHANISMS FOR IOMT ENABLED SHS

Security Focus	Authors	Strengths	Weakness
Secure data transmission	Mahender et al [6]	1) Any information concerning the identity and the patient's medical data was not revealed by the system. 2) The system achieved a better security level. 3) To save data transmission, the system was utilized.	It caused some major issues like large computation and also storage costs.
	Elhoseny et al. [7]	The system demonstrated its potential in effectively hiding the sensitive patient data confidentially to a transmitted cover image with higher undetectability and capacity but with minimum degradation in the acquired stego-image.	The system had provided satisfactory results, although, it demonstrate to work effectively.
Secure authentication	Rakesh et al. [8]	The hash variable value did not rely upon the hash functions for improving network security. The latency or delay of the system was not affected by the deviation in hashing.	A safe effective communication was not given by the system.
	Xu Cheng et al. [9]	The system built on community medical IoT system ensured the nodes' legality and communication security.	The system had a high computational expense and less security.
Confidentiality	Xuran Li et al. [10]	1) It protected the patient's confidential medical data amassed by means of medical sensors. 2) The eavesdropping risk was drastically reduced by the system.	It might cause partial perfect secrecy.
Access Control	Xunbao Wang et al. [11]	The system could be protected effectively during access process and transmission without loss in performance.	The manifold identities of the medical staff were not considered by the system.
privacy-preserving	Jing Wang et al. [12]	The system ensures data privacy during model training and also guarantees the security of the trained model.	The ML models are not supported by the system.

healthcare. This section summaries some state-of-the-art works as follows:

Faisal Alsubaei et al. [13] proposed a framework for security assessment of web-based IoMT. The framework recommends security features for IoMT employing ontological scenario-based approach. Also, it is employed to assess the protection and impediment of IoMT approaches. The proposed framework has demonstrated its potential in adapting (1) emerging new technologies and stakeholders; (2) compliance with standards; and (3) granularity. In general, system administrators are responsible for formulating security-related decisions. But the proposed framework opens avenues for all stakeholders in SHS to gain experience in cutting edge technologies related to the field of IoMT security. The system proved its efficacy with evaluation results in terms of all assessment attributes. But the employed assessment attributes were not easy to interpret by novice users like medical staff, patients who lack security and technical knowledge.

Muhammad Asif et al. [14] proposed a technique to ensure privacy of medical data especially against the threats emerging internally within SHS. The system allows access only for authorized users such as doctors and patients to communicate across the physical boundaries. The system had implemented authorization defining the permissions and roles merely for

medical staff. Further, the system enabled to remove any conflicts in access control models. Also, it guarantees to provide secure communication amongst doctors and patients in an efficient way. The system proved to outperform when compared with other related recent access control models in literature. However, the system does not facilitate to perform copy and move operations on directory resource.

Jinquan Zhang et al. [15] examined an encrypted storage model and a secure energy-efficient communication utilizing the benefits of rivests cipher 4 (RC4) for electronic health records (EHR) within IoMT enabled SHS. The system employed MedGreen authentication algorithm based on bilinear pair and elliptic curve for establishing secure communication. Also, the system utilized MedSecrecy algorithm that leverages Huffman compression ad RC4 for efficient data storage. The developed algorithm demonstrated to maintain the effectiveness of RC4 encryption and reduce the length of ciphertext data. Also it improved confidentiality, security and randomness. The simulation and analysis results proved that the system was energy-saving, secure and very effective for EHR. But, the system was not suitable to obtain more possible user information. In addition, Table I summarizes the state-of-the art works related to securing IoMT devices and applications in SHS.

III. Review on Lightweight Security Approaches for IoMT

Norah Alassaf et al. [16] proposed a lightweight cryptographic technique for IoMT enabled SHS applications. The contribution investigated the characteristics of SIMON cipher and employed it for IoMT enabled SHS applications for attaining performance as of a practical perspective. The system recommended to add an enhancement via original SIMON cryptography's implementation to diminish the computational complexity incurred owing to encryption. Also it enabled to preserve the practical balance between performance and security. However, the system did not give good results.

Zisang Xu et al. [17] introduced a key agreement and lightweight mutual authentication approach for IoMT. The system without employing symmetric encryption guaranteed to provide forward secrecy. The authors have utilized ProVerif software which is an automatic security verification tool to verify the system's security. The theoretical examination and experiential outcomes signified that the system drastically reduced the computational cost in comparison to the methods based on asymmetric encryption. Also, the system displayed lesser security risk compared to other lightweight approaches. Nonetheless, the system did not present the encryption and decryption time precisely.

Jianfei Sun et al. [18] proffered a lightweight fine-grained access control method to preserve the data privacy in IoMT enabled SHS. For the successful transformation of access policy and user attributes to the corresponding shorter length vectors, the system employed the optimized vector transformation process, whilst the other processes delivered longer and redundant vectors. This system was very significant in reducing the cost overhead incurred during decryption, key generation and encryption phases. Later, the system employed CP-ABE to gain the benefits of fine-grained access control and lightweight policy hiding to support SHS and handle the offline/online transformation process. Nonetheless, the system did not support traceability and attribute revocation.

Xiuqing Lu and Xiangguo Cheng [19] launched a lightweight data sharing approach with an aim to secure IoMT devices. First, the system ensured to provide authorized access and privacy over the shared data. Second, the system employed effective integrity verification when the user attempts to download the shared data. Doing so, the system enables to avoid false computational outcome or query. At last, the approach achieves lightweight in accomplishing the patients' and users' operation. The security analysis results confirmed that the system can enable to share data securely and effectively in IoMT enabled SHS as well its efficient in terms of computational cost. The system was not flexible towards possible attacks like tag forgery, reader impersonation and message eavesdropping.

Mahdi Fotouhi et al. [20] presented a lightweight 2-factor authentication approach to secure IoMT. The system was secured against different attacks. Furthermore, the system executes the formal and informal security evaluation. The system's security verification had been authenticated via the ProVerif. Moreover, the system had been simulated via the OPNET network simulator and compared with various other methods with regard to performance and security needs. The simulation comparisons and outcomes determined that the system had been appropriate and supported with added security attributes when compared to the relevant approaches. However, the system couldn't exhibit the precise security level.

Ran Ding et al. [21] introduced a lightweight PP identity-based authentication system for IoMT. The system performed data authenticators computation, data integrity verification on edge server. Also, edge server was used to system's computation overload and manage the third-party verification. The system achieved data privacy by enabling the patient to encrypt and transmit healthcare data to edge server. Also, the system uses cloud server to enhance the availability of the patient's data. At last, the performance and security evaluation are conducted to show the system potential. However, the system encompassed a storage overhead issue.

IV. Review on Blockchain Approaches for Security IoMT

Seyed Morteza et al. [22] presented an effective and secure method called "MedSBA" for storing medical data in SHS. The method is based on blockchain technology to ensure user privacy. Also, the system attempts to achieve fine-grained access control over patient data employing attribute based encryption (ABE) in compliance to the general data protections regulation. The system employed private blockchain to revoke the instant access which is very challenging in ABE. The security is proven via the formal design, whilst, the system's functionality had been proven using BAN logic. The efficiency of the system with regard to computational complexity and storage is demonstrated by simulating MedSBA's using OPNET software. Nonetheless, the system did not support the exchange of cryptocurrency between the data consumer organizations and the individuals for data sharing.

Ashutosh Sharma et al. [23] proffered a blockchain approach integrating the benefits of smart contracts for IoMT. The system examined the dimensions which the smart contract and decentralization could provide in IoMT. The IoMT devices are deployed in appropriate place to capture the data concerning the application needs. Also, these IoMT devices are pre-programmed to process and transmit the captured data. The efficiency of the system is demonstrated in comparison to other related techniques in terms of performance parameters such as average latency, average energy efficiency and average packet delivery ratio. However, the system encompassed less service quality; it did not function efficiently.

Neha Garg et al.[24] launched an authentication key agreements scheme based on blockchain for IoMT environment, termed as BAKMP-IoMT. It offered secured key management among the cloud servers, personal servers and the implantable medical gadgets. Further, the system provides secure access to sensitive healthcare data and ensures that it is accessed only by authorized users. This achieved by storing all the sensitive healthcare data into blockchain which is stored in cloud. Comprehensive formal security analysis has been conducted utilizing the extensively acknowledged automated tool, AVISPA to show the potential of the system against various types of possible attack. The comparative analysis results indicated the efficacy of the proposed method, BAKMP-IoMT over other existing approaches in terms of security requirements, communication and computation costs.

Jie Xu et al. [25] presented a PP scheme based on blockchain for large scale health data. The scheme encrypts the health data utilizing fine-grained access control. In specific, the user transaction are utilized for key management that can allow the users to add or revoke authorized doctors. Moreover, it avoid medical disputes as doctor diagnosis and IoT data cannot be tampered or deleted once stored to blockchain. Experiential and security evaluation outcomes confirmed that the system can well be applicable for SHS. However, the insider attacks are overlooked by the system.

V. Review on Authentication and Authorization Techniques for IoMT

Venkata P. Yanambaka et al. [26] suggested a lightweight and robust authentication based on physical unclonable function (PUF) for IoMT. This scheme does not stores any IoMT device related data on server memory. The system validation is performed utilizing a hybridized oscillator arbiter PUF. The amount of keys utilized for authentication was approximately 240 based on the PUF used during system validation. The authentication scheme being lightweight can be utilized in several designs for supporting the design's scalability and increasing its robustness. However, the system failed to ensure that the messages from server could be authenticated by the client.

Xu Cheng et al. [9] studied a secure identity authentication for community medical IoT. Here, the authors have utilized node security for system initialization. Next, the identity authentication was developed utilizing the benefits of mechanisms such as signature, session key symmetric encryption, elliptic curve encryption algorithm, secure two-way method. An effective community medical IoT node and an update mechanism that are secure and reliable to update the session and authentication keys is investigated. On the community medical IoT, the nodes' legality, along with the communication security had been ensured by these measures. The scheme was further appropriated for the community medical IoT's scene via the analysis, along with the analogy of experiential performance. However, the system had high computational costs and more power consumption.

Deebak et al. [27] suggested a mutual authentication scheme to secure SHS which centered on the IoMT. The system leveraged cloud to support emergency treatment for patients over internet communication from medical experts. The system ensured to secure the sensitive medical records and also maintained the patient anonymity. Further, it delivered an authentic signature for executing the secured transmission between the communication nodes. But, the system did not ensure to validate the access for services with regard to unforgability, undeniability and verifiability. Also, the system was insecure against confidentiality, forgery of health-report, non-repudiation and patient anonymity.

In [28], Sanaz Rahimi Moosavi et al. recommended a secure and effective architecture based on smart gateways for authentication and authorization architecture in IoMT-enabled SHS. Here, the smart gateways deployed in each healthcare sensors performed authentication and authorization and reduced the sensor overload while maintaining the all security requirements. Notably, the system relied on DTLS handshake protocol which is regarded as key solution for IoT security. The analysis results confirmed that the proposed architecture rendered better security compared to centralized delegation architecture. However, for the possible attacks, the system was not resilient.

Lone et al. [29] introduced a secure communication for medical applications utilizing ABE for authentication in HetNet. Here, health related data are secured utilizing ABE. This has not only helped to reduce the communication overhead but also has secures health data from intruders [30]. The entire security technique is implemented using high-level protocol specification language (HLPSL). The system codes are validated using automated tool, AVISPA. However, system provided less security and failed to work effectively.

Muhammad Tahir et al. [31] examined a framework for authentication and authorization mechanism which is lightweight to support blockchain-enabled IoT networks existing in health-informatics. Random numbers are utilized in the authentication process that was linked by conditional joint probability. This enabled the system to establish secure connection for the data acquisition amongst IoT devices. The authors utilized automated tools and simulator such as AVISPA and Cooja for system validation and evaluation, respectively. The system had provided strong mutual authenticity along with improved access control. It also decreases both the communication along with computational overhead cost when weighted against others as shown by the experiential outcomes. However, the system provided less efficient.

Yang Xin et al. [32] suggested a multimodal biometric identification scheme in the IoMT. An effective matching algorithm utilized by the system was based on secondary computation of the Fishers vector (FV). Further, the system

utilized three different biometric techniques like finger vein, fingerprint and face. These techniques are fused at feature level. Also, the system employed fake feature in the process of feature fusion which arises most frequently in practical scene. For decrementing the cause of the system's accuracy rate, and for increasing its robustness, the fake picture was removed. The designed framework had achieved an improved recognition rate as showcased by the experiential outcomes. It offered higher security whist analogized with unimodal biometric system that are extremely significant for an IoMT platform. But, the system had provided low accuracy values.

VI. Review on Privacy Preserving Approaches for IoMT

Maria et al. [33] proffered a PP approach for IoMT based on elliptic curve digital signature. By edge computing servers, privacy preservation in data transmitted as of IoMT to the cloud is done by this system. Especially, the captured health data was concealed from edge device and the identity of IoMT devices, namely, wearable or smart devices remained anonymous to cloud. As this solution is based on elliptic curves cryptography approach, its implementation on IoMT devices was feasible and affordable. Nevertheless, the computation and communication cost of the system found to be high.

Dong Zheng et al. [34] recommended an effective PP scheme for sharing medical data in IoT environment. The system supported data sharing leveraging the benefits of ABE. Further, the system utilized the attribute bloom filter removing the attribute matching function to maintain the confidentiality of attributes involved in the access control policy definition. The system utilized offline or online encryption technology in the phase of encryption to enhance the encryption's efficacy. A huge quantity of work ought to be done at the encryption stage before knowing the message. The cipher text could be produced quickly when the message was known. The analysis results demonstrated the potential of the scheme for sharing data in IoT environment. However, the scheme failed to verify and validate the cipher text that was stored over cloud.

Deebak et al. [35] proposed a PP protocol for securing SHS where attacker cannot imitate legal user to gain illegal access to the handheld smart card. The authors have used random-oracle model to perform formal and resource analysis to demonstrate the effectiveness of system security. Moreover, they have built a IoMT enabled SHS with top security feature which was revealed by its performance analysis. For analyzing, the network parameters based on the NS3 simulator, the experimentation analysis was executed. Regarding the throughput rates, packet delivery ratio, routing overhead along with end-to-end delay for the system, the collected results had shown superiority when analogized to other prevailing protocols.

Raylin Tso et al. [36] proffered a PP scheme for data communication through protected multi-party calculation in the HC cloud that are equipped with sensors. The system was based on the FairplayMP framework that enabled programmers to execute such protocols who were not specialist secure computation theory. Additionally, it was appropriate for distributed environments and it supported any numeral of participants. For example, to communicate with n-disparate data servers, each sensor node requires one single secret key to be stored in advance. But, the system was stored with three secret keys in advance in each sensor to communicate with three data servers despite the system offers low-level security.

S. Sheeba Rani et al. [37] presented an optimum users-based secure transmission of data within IoMT. The system employed Chinese Remainders Theorem to produce the cipher text copy according to the chosen number of users. Further the system utilized metaheuristic algorithm to choose the user in IoMT. Through simulation, the secure data performance was proved in terms of computation time, the energy price, etc., The outcomes confirmed that secure data could be effective whilst applied for ensuring security chances in IoT-based SHS but, low security was offered.

Alia Alabdulkarim et al.[38] put forward a privacy preserving single decision tree techniques aimed at clinical decisions-support systems to diagnosis the symptoms without disclosing the patients' data to disparate network attacks on IoT devices. For protecting users' data, homomorphic encryption cipher was utilized. Moreover, for avoiding one party as of decrypting the data of other parties, nonces were utilized as they would utilize the identical key pair. In addition, the system performed better than the Naïve Bayes algorithm by 46.46% which was revealed by the simulation outcomes. Additionally, for showing that it satisfies the attribute value's frequency, hospitals' dataset's privacy requirements, and effectively diagnoses the symptoms, the system was evaluated. However low-security services were possessed by the system.

Rihab Boussada et al. [39] examined a privacy preserving aware data transmission fort IoMT enabled SHS. User pseudonyms as public keys were defined by lightweight Identity-based encryption which was constructed on the elliptic curves discrete logarithm (ECDL) method. The contextual along with content privacy necessities are satisfied by the system. Regarding smart things limited resource nature, it was based on an identity-centered encryption scheme and specific communication scenario. A wide security examination was offered for validating the system and the performance analysis demonstrates its efficacy. However, for the e-HC emergency, the system was inappropriate.

Solihah Gull et al. [40] recommended a reversible data hiding approach based on dual image with large capacity for IoMT based networks. Initially, the Huffman encoding scheme was used to preprocess the captured secret data. A codebook of 'd' bits are generated after Huffman encoding to encode indices which are decimal values. For acquiring dual stenos-images, the indices' value are partition into two parts and embedded into two images that are similar to each other. Though very large payload was shown by the scheme, it proved to maintain the perceptual quality at high level. A noteworthy improvement was offered by the system and also computationally effective that made it to be utilized in the network of IoMT. However, for controlling the underflow and overflow issues, there wasn't an effective strategy.

Pei Huang et al. [41] presented a practical technique that could validate patients with the noisy signal of electrocardiogram (ECG) as well as offered disparate private protection concurrently. Regarding the present moving status, the scheme could identify the motions and adapted the algorithm. By offering indistinguishability, The ECG templates' privacy

was protected. The system's effectiveness was analyzed and validated over online datasets. For validating the system, pilot analysis on human subjects was conducted. However, for attacking, the system was not flexible.

Zhiwei Wang et al. [42] inspected an effective blind batch encryption scheme based on Computational Diffie-Hellmans assumption, which could be demonstrated as secure. For secure and privacy preserving medical services in SHC, the system utilized protocol. In the frame of six classic attacks, the system analyzed the protocol and executed prototype in the platform of Intel Edison. The experiments revealed that the system was effective for 'cheap' communication protocols along with resource-limited devices. For limited-storage devices, the system might require a heavy cost.

VII. Conclusion

When the networks are employed at large scales, the major concern is security. The major area focused in IoMT enables SHS is the patients' security and privacy. In this direction, authentication and authorization scheme play a very crucial role in ensuring eavesdropping the sensitive healthcare data and are considered as critical security requirements. Thus, there is great need for effective new solution that can render end-to-end data protection. From the review, it can be observed that several schemes are published for securing IoMT devices. Nonetheless, owing to the several constraints such as power, size, implantable and wearable, these smart devices do not have required resources for implementing the existing machine learning based security schemes. Therefore, to ensure the security, privacy and trust of these smart devices, require an efficient new solution that can meet all the security requirement and span across the design space of cyber. Also, the survey analysis reveals that ECC algorithm, lightweight authentication, and blockchain method are offering best security compared to conventional algorithms. Thus, to build power-efficient and sustainable IoMT enabled SHS, the upcoming research must focus on developing effectual lightweight intrusion detection systems to secure and safeguard IoMT enabled SHS.

Acknowledgment

The authors are very grateful to thank their Deanship of Scientific Research at Prince sattam bin abdulaziz university, Saudi Arabia for technical and financial support in publishing this work successfully.

References

[1] K. K. Karmakar, V. Varadharajan, U. Tupakula, S. Nepal, and C. Thapa, "Towards a security enhanced virtualised network infrastructure for internet of medical things (iomt)," in *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, pp. 257–261, IEEE, 2020.

[2] R. Somasundaram and M. Thirugnanam, "Review of security challenges in healthcare internet of things," *Wireless Networks*, pp. 1–7, 2020.

[3] G. Hatzivasilis, O. Soultatos, S. Ioannidis, C. Verikoukis, G. Demetriou, and C. Tsatsoulis, "Review of security and privacy for the internet of medical things (iomt)," in *2019 15th international conference on distributed computing in sensor systems (DCOSS)*, pp. 457–464, IEEE, 2019.

[4] B. A. Alqaralleh, S. N. Mohanty, D. Gupta, A. Khanna, K. Shankar, and T. Vaiyapuri, "Reliable multi-object tracking model using deep learning and energy efficient wireless multimedia sensor networks," *IEEE Access*, vol. 8, pp. 213426–213436, 2020.

[5] T. Vaiyapuri, V. S. Parvathy, V. Manikandan, N. Krishnaraj, D. Gupta, and K. Shankar, "A novel hybrid optimization for cluster-based routing protocol in information-centric wireless sensor networks for iot based mobile edge computing," *Wireless Personal Communications*, pp. 1–24, 2021.

[6] M. Kumar and S. Chand, "A secure and efficient cloud-centric internet-of-medical-things-enabled smart healthcare system with public verifiability," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10650–10659, 2020.

[7] M. Elhoseny, G. Ramírez-González, O. M. Abu-Elnasr, S. A. Shawkat, N. Arunkumar, and A. Farouk, "Secure medical data transmission model for iot-based healthcare systems," *Ieee Access*, vol. 6, pp. 20596–20608, 2018.

[8] R. K. Mahendran and P. Velusamy, "A secure fuzzy extractor based biometric key authentication scheme for body sensor network in internet of medical things," *Computer Communications*, vol. 153, pp. 545–552, 2020.

[9] X. Cheng, Z. Zhang, F. Chen, C. Zhao, T. Wang, H. Sun, and C. Huang, "Secure identity authentication of community medical internet of things," *IEEE Access*, vol. 7, pp. 115966–115977, 2019.

[10] X. Li, H.-N. Dai, Q. Wang, M. Imran, D. Li, and M. A. Imran, "Securing internet of medical things with friendly-jamming schemes," *Computer Communications*, 2020.

[11] X. Wang, F. Chen, H. Ye, J. Yang, J. Zhu, Z. Zhang, and Y. Huang, "Data transmission and access protection of community medical internet of things," *Journal of Sensors*, vol. 2017, 2017.

[12] J. Wang, L. Wu, H. Wang, K.-K. R. Choo, and D. He, "An efficient and privacy-preserving outsourced support vector machine training for internet of medical things," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 458–473, 2020.

[13] F. Alsubaei, A. Abuhussein, V. Shandilya, and S. Shiva, "Iomt-saf: Internet of medical things security assessment framework," *Internet of Things*, vol. 8, p. 100123, 2019.

[14] M. A. Habib, C. N. Faisal, S. Sarwar, M. A. Latif, F. Aadil, M. Ahmad, R. Ashraf, and M. Maqsood, "Privacy-based medical data protection against internal security threats in heterogeneous internet of medical things," *International Journal of Distributed Sensor Networks*, vol. 15, no. 9, p. 1550147719875653, 2019.

[15] J. Zhang, H. Liu, and L. Ni, "A secure energy-saving communication and encrypted storage model based on rc4 for ehr," *IEEE Access*, vol. 8, pp. 38995–39012, 2020.

[16] N. Alassaf, A. Gutub, S. A. Parah, and M. Al Ghamdi, "Enhancing speed of simon: a light-weight-cryptographic algorithm for iot applications," *Multimedia Tools and Applications*, vol. 78, no. 23, pp. 32633–32657, 2019.

[17] Z. Xu, C. Xu, W. Liang, J. Xu, and H. Chen, "A lightweight mutual authentication and key agreement scheme for medical internet of things," *IEEE Access*, vol. 7, pp. 53922–53931, 2019.

[18] J. Sun, H. Xiong, X. Liu, Y. Zhang, X. Nie, and R. H. Deng, "lightweight and privacy-aware fine-grained access control for iot-oriented smart health," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6566–6575, 2020.

[19] X. Lu and X. Cheng, "A secure and lightweight data sharing scheme for internet of medical things," *IEEE Access*, vol. 8, pp. 5022–5030, 2019.

[20] M. Fotouhi, M. Bayat, A. K. Das, H. A. N. Far, S. M. Pournaghi, and M. Doostari, "A lightweight and secure two-factor authentication scheme for wireless body area networks in health-care iot," *Computer Networks*, vol. 177, p. 107333, 2020.

[21] R. Ding, H. Zhong, J. Ma, X. Liu, and J. Ning, "Lightweight privacy-preserving identity-based verifiable iot-based health storage system," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8393–8405, 2019.

[22] S. M. Pournaghi, M. Bayat, and Y. Farjami, "Medsba: a novel and secure scheme to share medical data based on blockchain technology and attribute-based encryption," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–29, 2020.

[23] A. Sharma, R. Tomar, N. Chilamkurti, B.-G. Kim, *et al.*, "Blockchain based smart contracts for internet of medical things in e-healthcare," *Electronics*, vol. 9, no. 10, p. 1609, 2020.

[24] N. Garg, M. Wazid, A. K. Das, D. P. Singh, J. J. Rodrigues, and Y. Park, "Bakmp-iomt: Design of blockchain enabled authenticated key management protocol for internet of medical things deployment," *IEEE Access*, vol. 8, pp. 95956–95977, 2020.

[25] J. Xu, K. Xue, S. Li, H. Tian, J. Hong, P. Hong, and N. Yu, "Healthchain: A blockchain-based privacy preserving scheme for large-scale health data," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8770–8781, 2019.

[26] V. P. Yanambaka, S. P. Mohanty, E. Kougianos, and D. Puthal, "Pmsec: Physical unclonable function-based robust and lightweight authentication in the internet of medical things," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 3, pp. 388–397, 2019.

[27] B. Deebak and F. Al-Turjman, "Smart mutual authentication protocol for cloud based medical healthcare systems using internet of medical things," *IEEE Journal on Selected Areas in Communications*, 2020.

[28] S. R. Moosavi, T. N. Gia, A.-M. Rahmani, E. Nigussie, S. Virtanen, J. Isoaho, and H. Tenhunen, "Sea: a secure and efficient authentication and authorization architecture for iot-based healthcare using smart gateways," *Procedia Computer Science*, vol. 52, pp. 452–459, 2015.

[29] T. A. Lone, A. Rashid, S. Gupta, S. K. Gupta, D. S. Rao, M. Najim, A. Srivastava, A. Kumar, L. S. Umrao, and A. Singhal, "Securing communication by attribute-based authentication in hetnet used for medical applications," *Eurasip Journal on Wireless Communications and Networking*, vol. 2020, no. 1, pp. 1–21, 2020.

[30] T. Vaiyapuri and A. Binbusayyis, "Application of deep autoencoder as an one-class classifier for unsupervised network intrusion detection: a comparative evaluation," *PeerJ Computer Science*, vol. 6, p. e327, 2020.

[31] M. Tahir, M. Sardaraz, S. Muhammad, and M. Saud Khan, "A lightweight authentication and authorization framework for blockchain-enabled iot network in health-informatics," *Sustainability*, vol. 12, no. 17, p. 6960, 2020.

[32] Y. Xin, L. Kong, Z. Liu, C. Wang, H. Zhu, M. Gao, C. Zhao, and X. Xu, "Multimodal feature-level fusion for biometrics identification system on iomt platform," *IEEE Access*, vol. 6, pp. 21418–21426, 2018.

[33] M.-D. Cano and A. Cañavate-Sanchez, "Preserving data privacy in the internet of medical things using dual signature ecdsa," *Security and Communication Networks*, vol. 2020, 2020.

[34] D. Zheng, A. Wu, Y. Zhang, and Q. Zhao, "Efficient and privacy-preserving medical data sharing in internet of things with limited computing power," *IEEE Access*, vol. 6, pp. 28019–28027, 2018.

[35] B. D. Deebak, F. Al-Turjman, M. Aloqaily, and O. Alfandi, "An authentic-based privacy preservation protocol for smart e-healthcare systems in iot," *IEEE Access*, vol. 7, pp. 135632–135649, 2019.

[36] R. Tso, A. Alelaiwi, S. M. M. Rahman, M.-E. Wu, and M. S. Hossain, "Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud," *Journal of Signal Processing Systems*, vol. 89, no. 1, pp. 51–59, 2017.

[37] S. S. Rani, J. A. Alzubi, S. Lakshmanaprabu, D. Gupta, and R. Manikandan, "Optimal users based secure data transmission on the internet of healthcare things (ioht) with lightweight block ciphers," *Multimedia Tools and Applications*, pp. 1–20, 2019.

[38] A. Alabdulkarim, M. Al-Rodhaan, T. Ma, and Y. Tian, "Ppsdt: A novel privacy-preserving single decision tree algorithm for clinical decision-support systems using iot devices," *Sensors*, vol. 19, no. 1, p. 142, 2019.

[39] R. Boussada, B. Hamdane, M. E. Elhdhili, and L. A. Saidane, "Privacy-preserving aware data transmission for iot-based e-health," *Computer Networks*, vol. 162, p. 106866, 2019.

[40] S. Gull, S. A. Parah, and K. Muhammad, "Reversible data hiding exploiting huffman encoding with dual images for iomt based healthcare," *Computer Communications*, vol. 163, pp. 134–149, 2020.

[41] P. Huang, L. Guo, M. Li, and Y. Fang, "Practical privacy-preserving ecg-based authentication for iot-based healthcare," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 9200–9210, 2019.

[42] Z. Wang, "Blind batch encryption-based protocol for secure and privacy-preserving medical services in smart connected health," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9555–9562, 2019.

High Speed Single-Stage Face Detector using Depthwise Convolution and Receptive Fields

Rahul Yadav[1], Priyanka[2]
ECE Department[1,2]
DCR University of Science and Technology,
Murthal, Sonepat, India, 131039
ORCID ID: 0000-0003-2542-112X[1]

Priyanka Kacker[3]
Institute of Behavioural Science
National Forensic Sciences University,
Gandhinagar, Gujarat, India, 382007

***Abstract*—At present face detectors use a large Convolutional Neural Network (CNN) to achieve high detection performance, which is a widely used sub-area of artificial intelligence. These face detectors have a large number of parameters which reduces their detection speed dreadfully on a system with low computational resources. This is a challenging problem to achieve good performance and high detection speed with finite computational power. In this paper, we propose a single-stage end-to-end trained face detector to address this challenging problem. The computational cost is reduced by using depthwise convolution and swiftly reducing the size of an input image. The early layers of the model use CReLU (Concatenated Rectified Linear Unit) activations to preserve the information and generate better representative features of the input. Respective Field (RF) blocks used in the model improve the detection performance. The proposed model is of 1.7 Megabytes size, able to achieve 42 FPS (Frame Per Second) on CPU (i5-8330H) and 179 FPS on GPU (GTX1060). The model is evaluated on various benchmark datasets like WIDER FACE, PASCAL faces and AFW and archive good performance compared to other state of art methods.**

***Keywords*—Artificial intelligence; computer-vision; Convolutional Neural Network (CNN); face detector**

I. Introduction

Face detection is defined as the problem of detecting and localizing faces in a given image. It is a basic and long-standing problem of active research in computer vision. Applications such as face recognition, face tracking and face hallucination, use face detection as a primary and essential preprocessing step. Many practical systems for facial analysis, surveillance and bio-metric, requires fast and accurate face detection.

There are two challenging problems encountered in face detection. The first problem is of classifying faces with a large variety of facial appearances from a complex background. Second of detecting faces of different sizes at different positions in given images. The two problems are related to computational cost and speed of face detection. It is a challenging task to develop a face detector that creates a balance between two problems. Another problem is that is the boundary of an object is blurred by imaging systems also [1], [2].

Face detection methods can be broadly divided into two categories, traditional methods and CNN based methods. The traditional methods, are very fast but does not have good accuracy. These methods use hand-crafted features to train the classifiers. Viola-Jones [3] and Deformable Part Models (DPM) [4] are good examples of traditional methods which have good speed with decent accuracy. The performance of these detectors decreases in an unconstrained environment. This is mainly due to non-robust handcrafted features.

The CNN based methods can achieve high performances at cost of speed. This significant improvement in the accuracy of face detection diverted researchers attention towards CNN based face detectors. CNN models can achieve high performance by using a large number of convolutional layers, which are also responsible for the slow speed of the detector. For example, some recent high performing face detectors like DSFD [5], Pyramidbox [6] and Retinaface [7], use large CNN models like VGG-16 [8] and Resnet-152 [9]. These CNN models consist of a large number of parameters, for example, VGG-16 has 100 million parameters and Resnet-152 has 65 million parameters. CNN methods [5], [6], [7] are slow, hence not suited for many practical systems. Cascade CNN [10], [11] can be used to improve the detection speed. But these detectors suffer two limitations. First, each stage of the cascade is trained and optimized separately which make training difficult and also affect its performance. Second, the speed of the detector directly proportional to the number of faces in an image.

In this paper, a lightweight single-stage end-to-end trained face detector with fast speed and good accuracy is proposed. The proposed method can be divided into two networks, backbone network which extracts feature from input images and detection network which localize the faces. The backbone network uses depthwise separable convolution with large strides to swiftly reduces the dimension of input. Instead of using the max-pooling layer model as given in [12], the proposed model use depthwise separable convolution to reduce the size because it adds extra feature layers and hence provides better feature representation. CReLU [13] activation are used to preserve the information while reducing the size of the input using large strides in the proposed network. The detection network consists a Receptive Field (RF) blocks followed by depthwise convolution layers. A feature map from RF blocks is used for detection.

The main contribution of this paper can be summarized as follows: (1) Propose a new lightweight backbone design to overcome the drawbacks of previous methods. (2) The new lightweight face detection method is proposed by integrating the backbone network with an RF-based detection network for fast and accurate face detection. (3) The experiments performed on multiple benchmark datasets show proposed

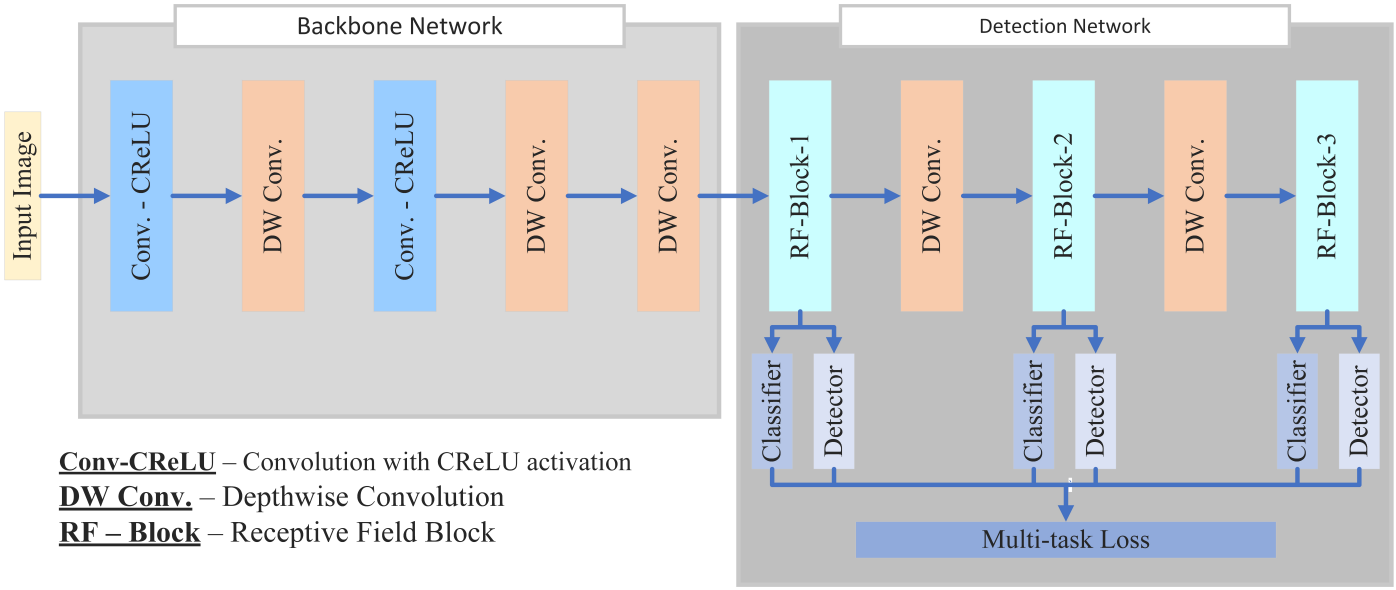


Fig. 1. General Frame Work of the Proposed Model. The Proposed Network can be divided into two parts i.e. Backbone Network and Detection Network.

method performs better than other methods. (4) Experiments performed on CPU and GPU hardware shows that the proposed method is suitable for practical systems. Hereafter the paper is organized as follows, Section 2 contain a brief review of available CNN based face detectors and techniques used in proposed methods. Section 3 is about the proposed method, it explains the framework of the method and its implementation details. Results obtained from experiments are discussed in Section 4, followed by the conclusion in Section 5.

II. Related Works

A. CNN based Face Detectors

Almost all modern days face detectors uses CNN architectures. The CNN based face detectors can be classified into three categories, i.e., cascade face detectors, region-based face detectors and single-stage face detectors.

The cascade face detectors divide the detection task into more than one CNN networks. CNN cascade structure introduced in [10], it consists of six CNN networks, three networks for each classification and calibration respectively. Architecture consisted classification network followed by a calibration network. MTCNN [11] reduced the number of networks to three by integrating classification and calibration task into one network. The first network is called P-Net, which proposes a facial region. Later two networks, O-Net and R-net, refines the proposals. The author in [14] divided P-Net into six sub-networks to detect faces at multiple scales. This improves detection performance for tiny faces. The detection performance of cascade face detectors are is improved by adding extra information about facial parts [15], [14]. In cascade framework, the first Network proposes the facial regions and subsequent networks process these regions. This makes the speed of detectors dependent on the number of faces in the images and it is a major limitation of these detectors.

The region-based and single-stage detectors are also known as two-stage and single-stage detectors, respectively. Both the detectors were developed for generic object detection. Later these detectors were modified to be used for face detection. The region-based detectors have two stages, first stage generates object proposal regions from proposal generators. The precise location and class of the object are estimated in the second stages. R-CNN based face detectors [16], [17] use RPN (Region Proposal Networks [18]). The performance of the method is further improved by CMS-RCNN [19] by adding contextual in formations. The region-based detectors use large CNN networks for the second stage. This lead to high detection accuracy but framework processing speed becomes slow.

Single-stage eliminates the region proposal stage and use a single stage to make predictions. These detectors are computationally efficient compared to region-based detectors but suffer detection accuracy. Single-stage face detectors are inspired by generic object detectors like YOLO [20] and SSD [21]. These detectors have attracted more researchers because of there high-speed detection. Different architectures [5], [6], [7] have been proposed recently. Lightweight CNN architecture [12], [22] uses inception module, CReLU activation and also propose densification strategy for anchors to improve recall. LFFD [23] paper proposes an anchor-free lightweight model by using Receptive Fields (RF) as natural anchors for detection. The model parameters were significantly reduced to 0.1 million in [24] by integrating the image pyramid with the CNN network and using weight sharing. But still, there is a large room for improving the processing speed without sacrificing detection accuracy.

B. Receptive Field (RF) and Dilation

Receptive Fields (RF) in CNN are inspired by the human visual system. RF in the visual system is neurons respond to a particular area of the retina. Similarly, in CNN each neuron has an RF field that responds to a particular area of an input [25]. In other words, RF defines the local region of an image to which the neuron will respond. The area RF is determined by the kernel size used in the convolution layer. RF has two important

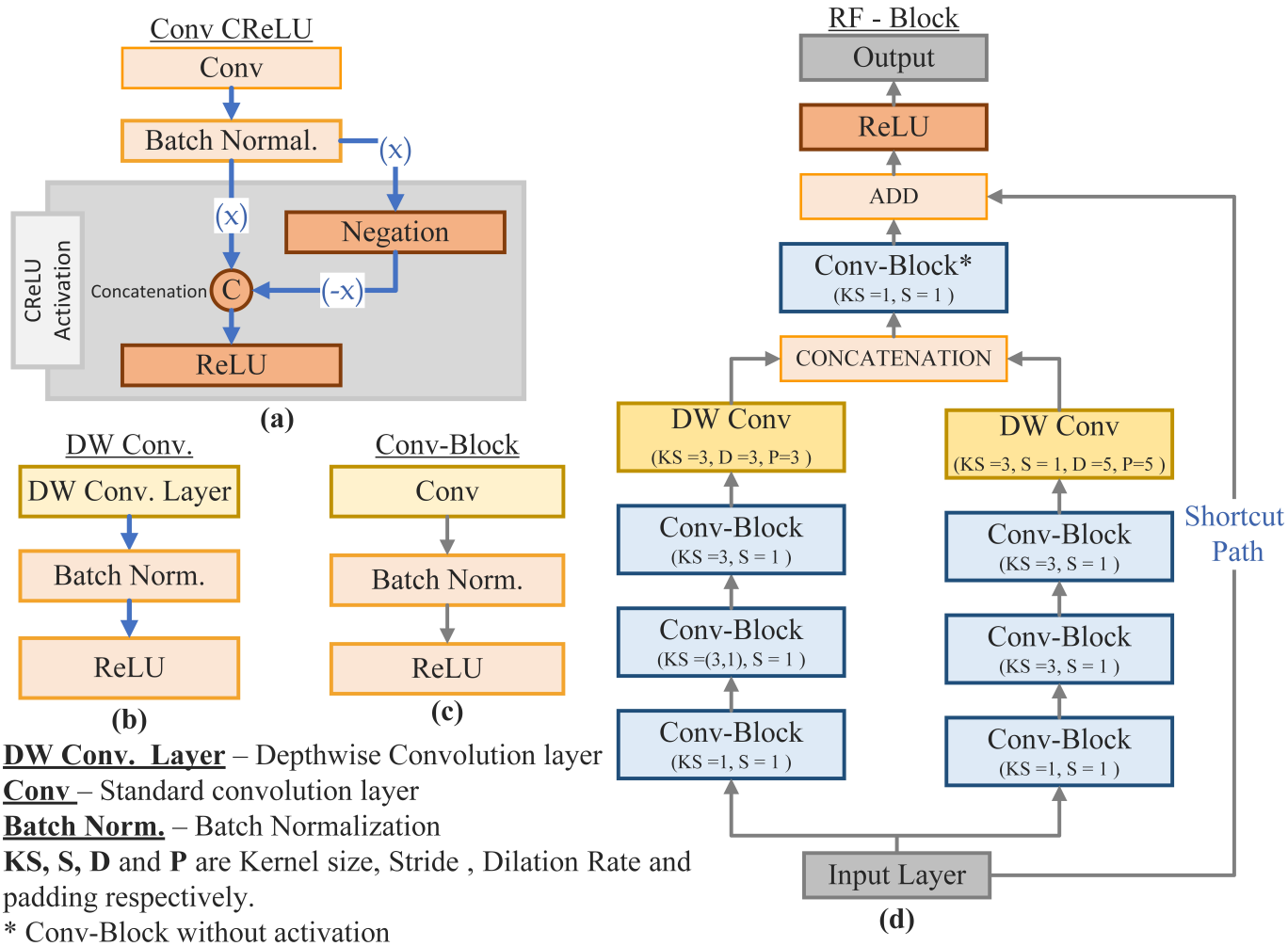


Fig. 2. Detailed Architectural Description of Blocks used in Proposed Methods. (a) Showing the Conv-Blocks with CReLU activation, (b) show Depthwise Separable Convolution used DW Conv block in Backbone Network and RF-Block, (c) Standard Convolution Layer for Conv-blocks used in Backbone Network and RF-Block and (d) detailed Architectural view of RF-Blocks used in Detection Network of Proposed Method.

properties, first, each neuron in CNN has unique activation for a given image region and second, pixels surrounding RF have a large impact on activation. The impact of neighbouring pixels can be represented as Gaussian-Distribution [23], and known as ERF (Effective RF), This RF also helps in detection by adding contextual information to the network.

The RF of CNN can be increased by adding convolution layers, depthwise convolution or by using dilated convolutions [26]. Adding convolution layers (increasing depth of networks) increases computational cost. So using dilation convolution and depthwise more effective way to increase RF. The dilation convolution introduced in [27] as astrous convolution. Dilation convolution is very similar to conventional convolution layer except there is a gap in kernel values which is decided by dilation rate. The author in [12], [23] used RF and dilation convolution for face detection.

C. Depthwise Separable CNN

Many states of art CNN architectures [28], [29] uses depthwise separable convolution layer. The depthwise convolution layer is computationally more efficient than a standard convolutional layer. The standard convolution layer performs convolution operation on input volume and combines generated features in one step. The computational cost of standard convolution is $D_k.D_k.M.N.D_F.D_F$ [28]. Where $D_F and D_k$ is the spatial dimension of input feature and kernel size respectively. While M, N are the number of channels in input features and number of convolution filters respectively. To reduce the computational cost, the one-step process is divided into two steps by using factorized convolution also known as depthwise separable convolution.

The first step is depthwise convolution operation performed on each channel of the input feature map separately. two assumptions are made in this step, (1) that the number convolution filter is equal to the number channels of the input feature map and (2) the spatial size of input and output feature maps are the same. If depthwise convolution is performed on input feature map of spatial size $D_k \times D_k \times M$ using filter of $D_F \times D_F$ spatial size. Then $D_k \times D_k \times M \times D_F \times D_F$ multiplication operations are performed in this stage.

Second step is point wise convolution, 1×1 convolution is performed across M channels output of depthwise convolution. This help to gain cross channel information and linearly combine the output. If N filters of 1×1 dimension is used on $D_k \times D_k \times M$ depth wise convolution output. Then $D_k \times D_k \times M \times N$ multiplication operations are performed. Therefore the computational cost of depthwise convolution is $D_k \times D_k \times M \times (D_F \times D_F + N)$. For qualitative comparison consider an image of $100 \times 100 \times 3$ is passed through depthwise convolution layer and standard convolution layer. If N = 10, then standard convolution performs 2.7×10^7 operations while dethwise convolution perform 5.7×10^6 operations which is approximately 4.7 times less than of standard convolution operations.

III. Proposed Method

In this section, the overall framework of the proposed model is introduced. Followed by a detailed description of model training.

A. Overall Framework of Proposed Model

Proposed face detectors can be divided into two networks, i.e. backbone network and detection network as shown in Fig. 1. The backbone network designed to swiftly reduce the dimensions of the input images without losing information during the process. The backbone network consists of a total of five convolution blocks, the first and third blocks are standard convolution layers with CReLU activation and having large strides. The remaining second, fourth and fifth blocks are depthwise separable convolution block. CReLU is used for its reconstruction property which is of information preserving nature, which leads to features reconstruction power of CNN [13]. CReLU activation is applied by concatenating the linear response of the CNN layer and its negation and passing it through ReLU activation as shown in Fig. 2(a). Mathematically it is defined as:

$$\forall x \in \mathbb{R}, \quad \rho_c \triangleq ([x]_+, [-x]_+) \tag{1}$$

Where $\rho_c : \mathbb{R} \rightarrow \mathbb{R}^2$, CeRLU activation and x is linear response of CNN network. From the above equation 1, it can be easily deduced that CReLU activation perverse both negative and positive response. Hence CReLU scheme produces representative features of input data [13]. To reduce the computational complexity depthwise separable convolution block are used. These blocks consist of depthwise convolution followed by batch normalization and ReLU activation as shown in Fig. 2(b). The feature obtained from a backbone network is feed into the detection network for further processing. The detection network is based on the cascade structure of SSD [21]. The model uses features from RF Blocks, which are spatially decreasing but have increasing respective field. Feature maps from different layer form multi-scale feature map to handle faces of variable sizes. RF-block-1, RF-block-2 and RF-block-3 are associated with anchor boxes to detect faces of small, medium and large sizes respectively. Multi-layer, multi-branch RF-blocks uses different kernels and dilation rates. This design has the advantage of classifying faces (with facial variation) from a complex background.

RF-blocks consist a bottleneck structure and residual connection as [30], [31]. The first layer of multi-branch design is 1×1 convolution, used to reduces the channel in feature maps. Then to reduce the computational cost 3×1 and 1×3 convolution is used. To increase the non-linearity and effective receptive field, depth-wise separable convolution with different dilation rate are used. Increased non-linearity, generates a more robust feature representation of the input. The increased effective receptive field helps to capture more contextual information for accurate classification. The branches are concatenated and a shortcut path is added to it. Fig. 2(d) shows the detailed architecture of the RF-block. Figure 2b and 2c show the architectural view of convolutional and depth wise separable convolution used in RF-blocks. In the model, each convolution layer is followed by batch normalization and ReLU activation respectively. This is done to reduce overfitting, induce sparsity and to handle the vanishing gradient problem.

B. Implementation Details

The model uses the anchor of 1:1 aspect ratio and densification strategy of [12]. The scale of anchors for RF block-1 are 32,64 and 128, for RF block-2 is 256 and RF block-3 is 512 pixels. The model is trained on WIDER FACE [32] training data set. This dataset consists of 12880 training images with different sizes, occlusion and blurriness levels. The training data is prepared by removing extremely small faces (height or width less than 15 pixels), heavily blur and occlude faces. For data augmentation, different strategies like random cropping, horizontal flipping, scale transformation and colour distortion are used during training. During training, the ground truth anchor boxes are matched to the predicted bounding box if the jaccrad index is more than 0.40. The multi-box loss objective function [21] is used in a training. It is a weighted sum of cross-entropy loss for bounding box confidence and smooth L-1 loss for bounding box coordinate regression. It is defined as:

$$L(c_i, d_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(c_i, c_i^*) + \frac{\lambda}{N_{reg}} \sum_i L_{reg}(d_i, d_i^*) \tag{2}$$

where, $L(c_i, d_i)$ is multi-box loss for given c_i confidence score of i^{th} bounding box with d_i coordinates. $L_{cls}(c_i, c_i^*)$ is cross entropy loss between predicted confidence score c_i^* bounding box and ground truth confidence score c_i. $L_{reg}(d_i, d_i^*$ is smooth L1 loss for predicted and ground bounding box coordinates. λ is hyper parameter used to balance the sum of losses ($\lambda = 2$ is used for training the network). Model is trained using batch size 32 for 280 thousand iterations. SGD optimizer used in training have 0.9 momentum, 5×10^{-4} weight decay. Model is trained using variable learning rates of 10^{-3}, 10^{-4} and 10^{-5} for 160K iterations 10^{-3}, 80K and 40K iterations respectively. The model is implemented using PyTorch framework[1].

IV. Results and Discussion

In this section proposed face detection algorithm is evaluated on the benchmark datasets, followed by speed comparison with available lightweight models.

[1]https://pytorch.org/

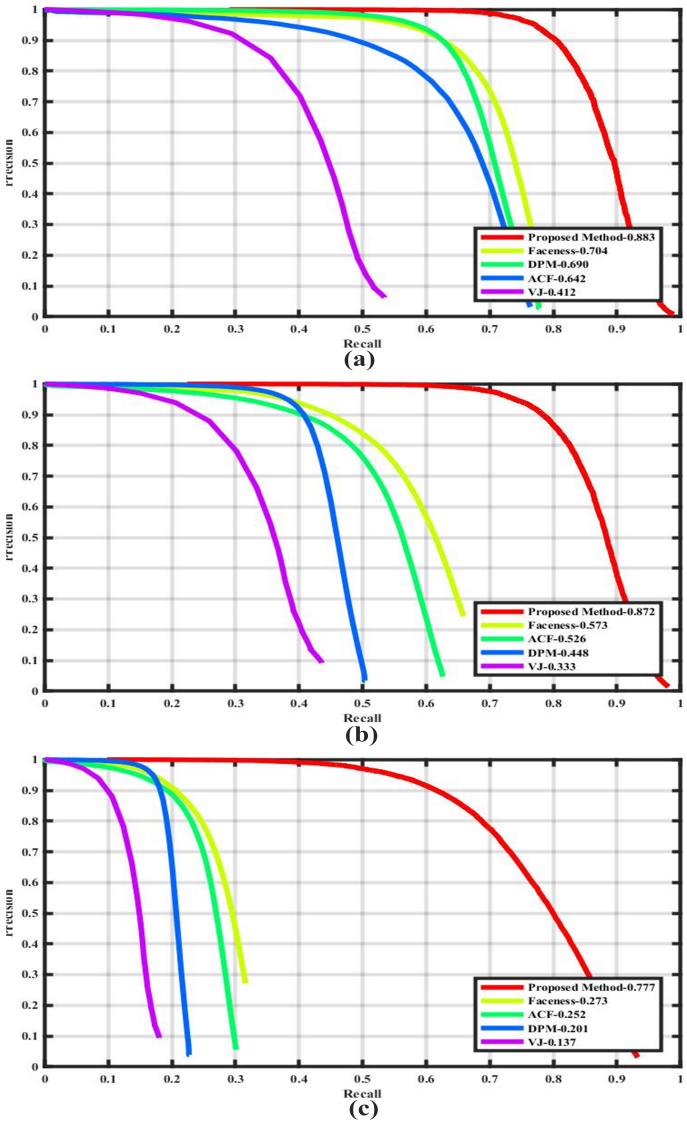


Fig. 3. PR Curve Comparing results of Proposed Methods and other Methods on (a) easy (b) medium and (c) hard validation subsets.

A. Experimental Setup

The proposed method is implemented using Pytorch version 1.6.0 on i5-8330H@2.30GHz processor system with 16 Gigabytes RAM and NVIDIA GTX 1060 GPU (Graphical Processing Unit).

B. Evalution on Benchmark Dataset

The proposed algorithm is evaluated on three benchmark face detection dataset, WIDER FACE [32], Pascal Face [33] and AFW [34] [33]. The proposed method is compared with other state of art lightweight detector using Average Precision (AP) percentage metric and PR (Precision-Recall) curves.

1) WIDER FACE Dataset: The dataset contains total 32203 images of faces different pose, scale, facial expressions and illumination. The dataset contain training and validation set. Validation set have three subsets validation data based on difficulties level of face detection, these are easy, medium and hard. The proposed method is trained on training set and validated results on all three validation subsets. Proposed method is validated against baseline methods [3], [4], [35], [36] and other methods [11], [12], [22], [37], [38], [39], [24]. Table I shows the results of performance comparison proposed methods with other methods. The proposed method shows the better result on easy and medium validation set, comparable result on hard dataset. This could be due to the fact that the network was trained on face which have height or width greater than 15 pixels and heavily occlude and blur faces were removed from the training set. Fig. 3 shows the PR curve of the proposed method compared against base line methods.

2) AFW and PASCAL Face Dataset: AFW dataset is Flickr images collection of 205 images with 473 face annotation. Table 2 shows the performance comparison of proposed method with standard methods using mAP% metrics. The

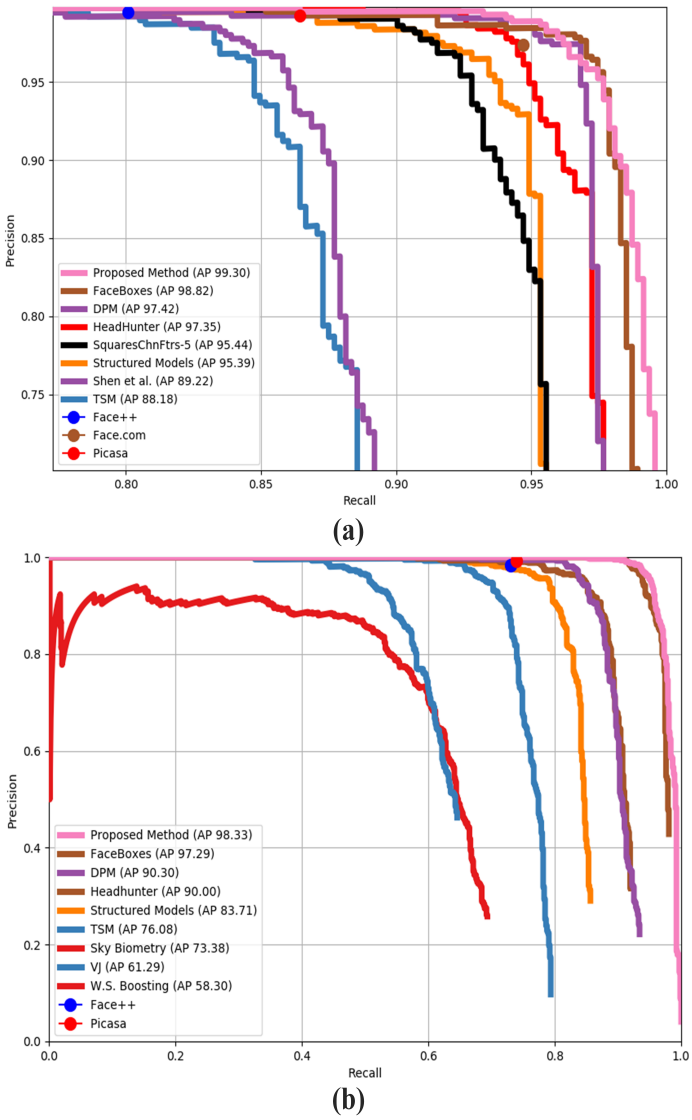


Fig. 4. PR Curve Comparing results of Proposed Methods and other Methods on (a) AFW and (b) Pascal dataset.

TABLE I. PERFORMANCE RESULT OF PROPOSED METHOD ON THE VALIDATION UBSETS OF WIDER FACE. THE REPORTED VALUES ARE MAP%

Methods	Easy	Medium	Hard
VJ*[3]	41.2	33.3	13.7
DPM* [4]	69.0	44.8	20.1
ACF* [35]	64.2	52.6	25.2
Faceness* [36]	70.4	57.3	27.3
MTCNN [11]	85.1	82.0	60.7
Faceboxes [12]	79.1	79.4	71.5
Faceboxes-2 [22]	87.9	85.7	77.1
ICC-CNN [37]	85.1	82.9	77.2
FDCNN [38]	73.3	67.8	51.0
Fastfaces [38]	83.3	79.6	60.3
Luo *et. al.* [40]	87.1	87.3	78.0
Proposed method	88.30	87.2	77.7

proposed method shows the better performance then other methods. [4], [12], [33], [41], [42]. Fig. 4(a) shows PR curve for proposed method,standard methods and commercial face detectors(Face.com, Face++ and Picasa).

Pascal Face dataset is formed from pascal person layout dataset. It contain 851 images with 1335 face annotations. The comparison of proposed method with standard dataset [3], [4], [12], [33], [42], [34] is given in Table II. Fig. 4(b) shows the PR curve of proposed method, standard method and commercial methods. Proposed method showed better results on dataset.

TABLE II. PERFORMANCE COMPARISON PROPOSED METHOD WITH OTHER METHODS ON AFW AND PASCAL DATASET. THE REPORTED VALUES ARE MAP%

Methods	AFW	PASCAL face
VJ [3]	-	61.29
DPM [4]	97.42	90.30
Headhunter [4]	97.35	90.0
SquareChnFtrs [4]	95.44	-
StrusctredModel [33]	95.39	83.71
TSM [42]	88.18	76.08
Shen *et al.* [41]	89.22	-
WSBoosting [34]	-	58.30
Faceboxes[12]	97.29	98.82
Proposed method	99.30	98.33

C. Running Efficiency

To check the practicality of a proposed method, it is tested on CPU and GPU hardware. Results obtained are then compared against the state of art methods reported running efficiencies in original paper.

TABLE III. RUNNING EFFICIENCY COMPARISON OF THE PROPOSED METHOD WITH STATE OF ART METHODS. THE SPEED OF THE METHOD ON CPU AND GPU ARE REPORTED IN FPS (FRAME PER SECOND)

Method	CPU	GPU
Faceness [36]	–	20 (Titan Black)
MTCNN [11]	16 (i7-4770K)	99(TitanX)
Faceboxes [12]	20 (E5-2660v3)	120(TitanX)
Faceboxes-2 [22]	28 (E5-2660v3)	245(TitanX)
ICC-CNN [37]	12 (i7-4770K)	40 (Titan)
FDCNN [38]	–	31 (GTX 1080)
ACF [35]	20 (i7-3770)	–
Luo *et. al.* [40]	50 (i7-6850 K)	180 (RTX 2080Ti)
Proposed method	42 (i5-8330H)	179 (GTX 1060)

The qualitative results are summarized in Table III. The results compared on image size 640X480. The detailed description of system on which test was performed is mentioned in section above. The proposed methods performance is very satisfactory, but it has comparatively slow then [40] because hardware (both CPU and GPU) on which the experiments performed are computationally inferior to other hardware on which the state of arts methods are tested.

V. CONCLUSION

This paper introduces a fast and high performing face detector. The high processing speed is achieved by using a lightweight backbone network. The feature extractor rapidly reduces the size of input without losing information during this process. The information is retained by the CReLU activation function. The performance of the face detector is achieved by efficiently utilizing the feature maps obtained from the feature extractor. The detector having RF blocks imitating the human visual system. As the results suggest the proposed method works well on the images with images having faces of the height of more than 15 pixels. The model limitation to detect tiny faces and heavily occluded faces. The proposed model can further be compressed using CNN optimization techniques such as pruning. The experiments performed using proposed face detectors on benchmark datasets has shown good results and have high processing speeds on both CPU and GPU devices.

ACKNOWLEDGMENT

The authors wish to acknowledge the National Project Implementation Unit (NPIU), a unit of the Ministry of Human Resource Development, Government of India, for the financial assistant-ship through the TEQIP-III Project at Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Haryana. The authors would like to thank the editors and anonymous reviewers for providing insightful suggestions and comments to improve the quality of the research paper.

REFERENCES

[1] P. Kaur, I. Lamba, and A. Gosain, "A robust method for image segmentation of noisy digital images," in *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*. IEEE, 2011, pp. 1656–1663.

[2] P. Kaur, A. Soni, and A. Gosain, "Image segmentation of noisy digital images using extended fuzzy c–means clustering algorithm," *International journal of computer applications in technology*, vol. 47, no. 2-3, pp. 198–205, 2013.

[3] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[4] M. Mathias, R. Benenson, M. Pedersoli, and L. V. Gool, "Face detection without bells and whistles," in *European conference on computer vision*, 2014, pp. 720–735.

[5] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: dual shot face detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[6] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 797–813.

[7] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5325–5334.

[11] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[12] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Faceboxes: A CPU real-time face detector with high accuracy," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 1–9.

[13] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *international conference on machine learning*, 2016, pp. 2217–2225.

[14] D. Zeng, F. Zhao, S. Ge, and W. Shen, "Fast cascade face detection with pyramid network," *Pattern Recognition Letters*, vol. 119, pp. 180–186, 2019.

[15] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, and W. Liu, "Detecting faces using inside cascaded contextual cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3171–3179.

[16] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 650–657.

[17] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[19] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection," in *Deep learning for biometrics*. Springer, 2017, pp. 57–79.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016, pp. 21–37.

[22] S. Zhang, X. Wang, Z. Lei, and S. Z. Li, "Faceboxes: A CPU real-time and accurate unconstrained face detector," *Neurocomputing*, vol. 364, pp. 297–309, 2019.

[23] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan, "Lffd: A light and fast face detector for edge devices," *arXiv preprint arXiv:1904.10633*, 2019.

[24] J. Luo, J. Liu, J. Lin, and Z. Wang, "A lightweight face detector by integrating the convolutional neural network with the image pyramid," *Pattern Recognition Letters*, 2020.

[25] G. Kobayashi and H. Shouno, "Interpretation of resnet by visualization of preferred stimulus in receptive fields," *ArXiv*, vol. abs/2006.01645, 2020.

[26] N. Adaloglou, "Understanding the receptive field of deep convolutional networks," *https://theaisummer.com/*, 2020 (accessed: 25.05.2020). [Online]. Available: https://theaisummer.com/receptive-field/

[27] L.-C. C. Author, G. Papandreou, I. Kokkinos, K. Murphy, and A. LYuille, "Deeplab: Semantic image segmentation with deep convolutional nets, astrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00474

[29] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[30] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[31] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 385–400.

[32] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[33] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014.

[34] Z. Kalal, J. Matas, and K. Mikolajczyk, "Weighted Sampling for Large-Scale Boosting." in *BMVC*, 2008, pp. 1–10.

[35] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *IEEE international joint conference on biometrics*, 2014, pp. 1–8.

[36] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3676–3684.

[37] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, and W. Liu, "Detecting faces using inside cascaded contextual cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3171–3179.

[38] D. Triantafyllidou, P. Nousi, and A. Tefas, "Fast deep convolutional face detection in the wild exploiting hard sample mining," *Big data research*, vol. 11, pp. 65–76, 2018.

[39] H. Zhang, X. Wang, J. Zhu, and C.-C. J. Kuo, "Fast face detection on mobile devices by leveraging global and local facial characteristics," *Signal Processing: Image Communication*, vol. 78, pp. 1–8, 2019.

[40] J. Luo, J. Liu, J. Lin, and Z. Wang, "A lightweight face detector by integrating the convolutional neural network with the image pyramid," *Pattern Recognition Letters*, 2020.

[41] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and aligning faces by image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3460–3467.

[42] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE conference on computer vision and pattern recognition*, 2012, pp. 2879–2886.

A Novel Framework for Modelling Wheelchairs under the Realm of Internet-of-Things

Sameer Ahmad Bhat[1], Muneer Ahmad Dar[2], Hazem Elalfy[3], Mohammed Abdul Matheen[4], Saadiya Shah[5]

Dept. of Electrical and Electronics Eng., Kuwait College of Science and Technology, Doha, 93004, Kuwait[1]
Dept. of Computer Science, National Institute of Electronics and Information Technology, Srinagar, India[2,5]
Dept. of Computer Science and Eng., Kuwait College of Science and Technology, Doha, 93004, Kuwait[3]
Dept. of Eng. Mathematics and Physics, Faculty of Eng., Alexandria University, Alexandria, 21544, Egypt[3]
Dept. of Computer Science, Common First Year, King Saud University, Riyadh, Saudi Arabia[4]

***Abstract*—Innovations in research labs are driven to global markets by applied, established standard engineering practices, using state-of-the-art research that most likely results in manufacturing highly effective and efficient engineered products. As a technology, that enables assistance to physically challenged people, wheelchairs have attracted researchers across the globe whilst showcasing an increased demand for higher production. However, wheelchairs in relation with the environments implementing Internet-of-things (IoT) devices, have been mostly overlooked to include the assessment of global market trends. Therefore, this paper proposes Acceptability Engineering (AE) framework to enhance the growth and expansion of markets relying on the environments, wherein wheelchairs can coordinate with IoT to enable smart technologies. AE as a standard engineering approach would help in – evaluating the characteristics of IoT-wheelchair environments, analysing their market trends, and highlighting the deficiencies between early and prevailing markets. This will significantly impact the manufacturers, who market wheelchairs specific to the IoT environments, and in addition manufacturers would be able to identify the potential users of their manufactured products.**

***Keywords*—Wheelchairs IoT; acceptability engineering; human-centered engineering; innovative technologies; early adopters**

I. Introduction

In recent years, research has focused on analyzing various existing models of wheelchairs, so as to transform them into highly advanced innovative products to enable human assistance. Wheelchairs embed intelligent systems which enable functional control of sophisticated drive control mechanisms, and recent studies are observed to focus on improving the overall control system, drive configuration mechanism and human-machine interaction [1], [2], [3]. Besides enabling assistance to the people with locomotion needs, wheelchairs also offer support in several other application areas, wherein they operate autonomously to accomplish certain time bound critical tasks. For instance, in On-Demand Transportation in Hospital Environment [4], Open Area Path Finding [5], wheelchair users in smart city planning [6].

In today's techno-savvy world, a number of things connect to the internet to transfer or update current status information, send or receive files, and reflect changes in databases. Objects around us when connected to each other through the internet are termed as Internet of Things (IoT) [7], in which every object holds a unique identity and can be accessed from anywhere anytime, through its exposed interfaces for monitoring and control, and acquire current status information. The number of objects connecting to the internet and the object-to-object communication rate, both are growing at an extraordinary rate. Billions and trillions of context aware, self-motivating remotely connected devices, integrate platforms for social integration and commercial business processes, for instance, social internet-of-things [8], Object with self-healing properties in IoT [7], Edge System for Smart Healthcare [9], ubiquitous computing [10], and ambient Intelligence [11], are predicted to widely expand the horizon of IoT. Presently, many IoT devices support home and business work flows, thereby ameliorate life experiences of people, as well as elevate the global market business trends. While IoT continues to dominate global markets, research in exploratory technology innovation processes specific to wheelchairs faces several challenges [12], [13], [14], and to overcome those challenges, whilst proposing concoct solutions to such problems, technology innovations in wheelchairs essentially require appropriate engineering methods.

Acceptability Engineering (AE) [15], a standard engineering approach, aims to build, estimate, and measure the effectiveness of innovative technology and users' acceptance relationships. As a systematic approach to assessment, AE reveals and overcomes the differences between early and late adopters, and between early and conventional markets. In order to determine the characteristics of technology innovation and user acceptance, AE provides a logical path for systematically analyzing such characteristics, and also helps in estimating market trends. AE primarily helps to access innovation technologies and their acceptance based on the characteristics defined by a user-centric engineering design, perception of the state-of-the-art technologies, users' realization, users' perception of information technology, and first-time user of an innovation technology.

In this paper, we aim to address the challenges that emerge as a result the two overlapping technological domains – the wheelchair technology domain and IoT domain, and we attempt to provide a feasible solution covering both domains and associated challenges. The main contributions of this paper are highlighted as under:

- Our proposed framework provides a common systematic design and assessment framework specific to the engineering that incorporates wheelchairs and IoT.
- The proposed framework defines key assessment strategies that can be applied in the design engineering process of manufacturing wheelchairs, specifically meant to operate in IoT networks.

- The proposed framework can help in accessing the role of assistive technology in relation with the IoT, as well evaluating users' perception of wheelchair under IoT domain.
- Our proposed framework can explain market trends to unleash the potential growth of wheelchair markets that operate under IoT networked environments, and would enable assessment of user markets, in particular users' adoption of assistive technology.

This study is organized into four different sections. The following Section II, provides a detailed background to the study. Section III describes the current trends in the context of wheelchairs. Section IV comprises a discussion on the Acceptability Engineering and IoT-Wheelchairs. Finally, Section V provides an overall conclusion of the study.

II. Background

Currently different types of wheelchair models are found in market and are typically used in a variety of social order fields. Patients with movement disabilities, in essence, have to perform real life tasks. Most of the tasks may require a person to move on from one location to the other, and wheelchairs act as a primary motion assistance tool for the patients with locomotive disabilities. Wheelchairs fundamentally vary in hardware and software, but the major difference that categorizes them is the propelling mechanism. Wheelchairs are basically operated either by a rider whirling the wheels directly, or may inevitably require someone to push the wheelchairs from behind, however, there also exist electronic wheelchairs that are propelled by powerful electric motors for navigation and other tasks, and these are typically controlled by a microcomputer-based circuitry [16]. Patients, who are weak enough to drive wheelchairs with their muscular arm force, make use of electronic wheelchairs. These type of wheelchair designs particularly target groups of people with walking disabilities, and allow digital control of the wheelchair, which may constitute an embedded processor for directional navigation system. The navigation system is typically controlled with a joystick that acts as a main controlling device. Electric wheelchairs mounted with sensors, actuators, and embedded processing elements, are categorized as Smart wheelchairs [17]. Most of them provide some additional controls for speed, navigation, breaks, alarms, etc. and due to the rigid and metallic construction, and digitally controlled electronic subsystems, they serve as an efficient hardware resource for building autonomous/semi-autonomous navigation robots [18], step climbing wheelchairs [19], brain- controlled wheelchairs [20], gesture-controlled wheelchairs [3] and other types.

The birth of IoT has led to its application in almost every aspect of human life. Recent studies reveal rapid updates in IoT technology, which implies massive scaling and outsourcing of the IoT products in the upcoming years [21]. However, looking at wheelchairs from the perspective of IoT environments, wheelchairs can be thought of as unilateral entities, disassociated and disconnected from IoT networks, and are unexpected to work in collaborative environments. Eventually, wheelchairs do require a space in IoT environments, wherein they can share, coordinate, communicate system/user status, and location information. This motivates us to employ and merge wheelchairs, as disconnected entities, within the IoT environments, thereby leading researchers to towards the newly emerging challenges. So far, several research studies have attempted look into the scenario of wheelchair operating in association with IoT networks, and this implies more innovations are yet to come. Therefore, analyzing the development process, market growth and user acceptance of wheelchairs in the IoT based environments needs focus, and is subject to a new approach of engineering.

As IoT is an emerging field, and is still in its early phase, exploration of wheelchairs networked within IoT environments, will present imminent challenging tasks with varying complex levels. Therefore, researchers will be able to explore existing gaps in knowledge, underlying within the composite domain formed by IoT layer and the wheelchair. From our literature survey, we observe that till date there is no engineering approach that specifically defines an evaluation methodology for the wheelchairs that would operate in IoT environments. Therefore, in this study, we propose Acceptability Engineering (AE) as an engineering method to model the framework that amalgamates the domains of assistive technologies and the IoT. The AE framework can support in developing, perceiving, and evaluating the role of innovative technology as well as highlighting the users' perception of assistive technology, in particular wheelchairs. AE can uncover the gap that prevails between the novice and later users of assistive technology, abreast to enabling a systematic track of assessment of emerging trends between initial and later markets. To explain the technology innovation, and its acceptance by the end users, AE can help in the design engineering process of assistive technology, that operates under the umbrella of IoT, that bases on human centered approach of engineering, users' perception of innovative technology and its acceptance, users' experience of applied information technology, and the evaluation of early adopters needs of engineered assistive technology.

Therefore, the paper guides through a logical perspective of how the application of AE can assist in the evaluation of innovation processes that drive wheelchair technology in the IoT based environments, how the application of AE in the innovation process will enable researchers to evaluate the system characteristics, which include, engineering the system, identifying interdisciplinary domains, determining user skill levels, technology innovation under consideration and potential adopters of technology in the markets.

III. Current Trends

The role of technology is increasing rapidly in improving mobility of physically challenged people. In order to understand the different types of human assisting technologies, specifically meant for people with disabilities, we need to setup benchmarks to compare and analyze the current technologies aimed at delivering human mobility. Using international standard assessments of World Technology Evaluation Center, the National Science Foundation began a study, which was conducted by a team of experts, in charge of collecting information related to current trends in technology, to perform mobility conversion for the people with motion disabilities [22]. Despite the limited scope of their study, the team highlighted seven mobility tasks, necessary for the development of mobility assistance. These seven tasks relate to – positioning, stability and relocation, operation, mobility, stir uphill, other motion tasks, and the transportation. While considering technology innovation for disabilities, those seven mobility tasks could serve as starting initial checklist, to assess and thoroughly define wheelchair devices. While considering technology innovation for disabilities, those seven mobility tasks could serve as starting initial checklist to assess and thoroughly define wheelchair devices. Prior research studies have attempted to

address and unpin only some of the above-mentioned tasks, for example, intelligent wheelchair for tennis [23], development of simulator and analysis tool for navigating in indoor locations [24], adaptive motion control for semi-autonomous wheelchairs [25], hands free control for wheelchair [26] and Voice Controllable Wheelchair [27]. In the last five years, there is an elevated trend to integrate intelligence systems into wheelchair control boards [28] to enable automatic or semi-automatic wheelchairs design. The recent innovative solution proposed and developed by AT&T for the wheelchair employ IoT, enable access to data that influence everyday lives of wheelchair users [29]. Through the remote access to this global data, stored on highly secured cloud platforms, allows data sharing among various stake holders of the system, to improve system usage and benefit the users' as well as designers.

With the Smart Technology dominating all the areas of technology nowadays, the checklist of these seven parameters is inadequate and needs extension, and this extension is further subject to the wheelchair application domains; those domains where the assistive technology is deployed for serving humans in a variety of areas other than the area of typical human mobility assistance. Some areas may necessitate inclusion and careful estimation and evaluation of all seven parameters, whereas others may need only a few. For example, if a wheelchair is employed for an individual with minor motion disability, it does not require critical monitoring, evaluation and inclusion of smart technologies in the wheelchair design. On the other hand, some wheelchair designs may require critical evaluation of above-mentioned parameters. For example, people with serious locomotive problems need time bound, frequent systematic and careful monitoring, and there is no room for failures as it may result in loss of human lives. Consequently, regular monitoring and safety measurements remain the top priorities in such systems. Most applications require mounting sensors on wheelchairs, to record each and every critical activity, happening in either real or non-real times. Real time activity monitoring in wheelchairs requires addition of the other dimensions, like network connectivity, data and information security, data storage, measurement and estimation of various hardware and software parameters, to the previous checklist of seven parameters. Monitoring and tracking wheelchair status round the clock is a difficult task as it requires someone to be always available with the wheelchair. Also, remote assistance is often preferred to control devices that are out of reach. These issues introduce us to the concept of internet of things, in which each object is viewed as a networked entity to provide activity data and information from anywhere and anytime. By analyzing these recent trends, we can briefly state the current needs, that form the core basics and are prerequisite to the design and development of wheelchairs. Following are specified trends, which serve as a guide for research scholars working towards the future research agendas in the current context.

- Modelling and development of efficient wheelchairs designs through state-of-the-art innovative technologies.
- Technology transformation and quality production, involving multi-disciplinary approach to develop highly structured innovative products.
- Understanding the recent realm of current innovative technologies and their applications.
- Devising frameworks to measure behavioral satisfaction levels of various wheelchair systems users.
- Identifying potential adopters of newly innovative technologies.

IV. Acceptability Engineering and IoT

To underline IoT-Wheelchair design characteristics, this study is based on frameworks primarily focusing on users' innovative technology acceptance. As discussed in the previous sections, Acceptability Engineering (AE) incarcerates the prevailing mutual relationship between innovative technology and user acceptance. It can be used to evaluate current trends and characterize wheelchair technology in relation with IoT. Fig. 1 shows the proposed framework for Wheelchair and IoT co-domain, and the following sub-sections next, describe the framework in more details.

A. Perception of Human-Centered Design Engineering

AE mainly concerns research on human-centered innovative technologies, the association between innovative technologies and user acceptance. It systematically examines users' technology acceptance behavior through cognitive, emotional and social factors. IoT-Wheelchair design more importantly concerns users' acceptance. Therefore, it is quite satisfactory to apply AE's human centric approach in the design process. The Human-centered engineering approach also called as user centered design (USD), directly places a user at the core of the design and decision-making engineering process. Human centered design approach engaged in AE defines how users accept innovative technologies, in contrast to building advance technology systems. Therefore, development of well-organized and robust products, is a natural outcome of this approach, and this will help us in – estimating factors responsible for higher user acceptance rates, designing acceptable alternatives to failing innovative technologies, evaluation of user acceptance and prediction of technology innovation success rates.

The cognitive and organizational factors are an essential construct in the product design and safety assessment processes. Accuracy level of procedure designs and interfaces can be highly improved via the human machine interaction analysis. This places responsibilities on cognitive science engineers to precisely assess the psychological behavior of IoT-wheelchair users. The primary and transient conditions observed IoT-wheelchairs, can be compared and examined through various methods and interfaces. The theories of cognition form the basis to the Human Factor Analysis Methods and are applicable to several engineering disciplines. These can be applied within four different areas of engineering: system designing, measuring safety levels, training, and accidental analysis.

Safety, is the next and important step in design process for wheelchair users, as well as the wheelchair itself. Studying IoT-wheelchair relationship, poses certain questions to the engineers. The first question to be answered is how to ensure the real time data communication. Next is to assess the system behavior based on some prior Safety Instructions Model. To support a driving wheelchair, safety driving assistance system can be setup through the IoT. As IoT offers opportunities to a number of users to synchronize and share their experiences through interconnection, it guides engineers to develop a framework for automatically adjusting the safety parameters, including context awareness. These safety parameters could be specified in the IoT model or could be employed directly in the software models of wheelchair systems. Therefore, AE can support in developing safety models based on user acceptance levels, in the context of IoT and wheelchairs.

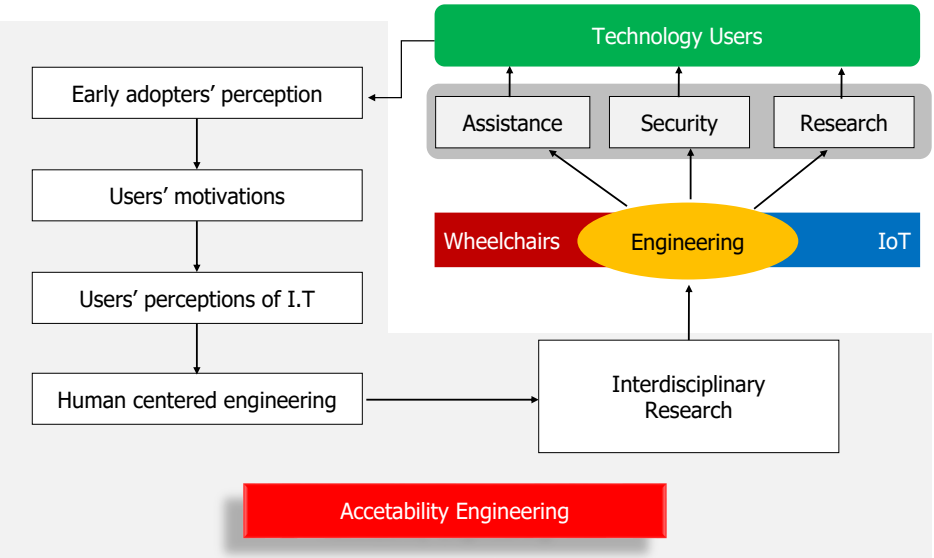


Fig. 1. Acceptability Engineering for Wheelchairs and IoT Domain.

IoT-wheelchair design and development requires it to be user-centric, with all of its components being highly natural. During our literature search, we realized that only a small number of studies have followed up with this kind of approach. Therefore, it needs to be mentioned that designing wheelchairs using human-centered design would prove to be a highly significant methodology, as its main objective is to assist and satisfy user requirements.

B. Technology Perception using Interdisciplinary Research

Wheelchair design engineers rigorously research for developing efficient hardware and software, and activities demand engineering from multiple disciplines to work collaboratively. Often, engineering teams work so closely in association, that logical domain boundaries seem overridden. This transdisciplinary approach in the design engineering process looks a singular domain with blurred boundaries between the unlike domains. To solve complex problems, the trans-disciplinary approach proves to be the best choice. Consequently, the IoT-wheelchair relation requires extensive artwork of engineers to transmute wheelchair model into an IoT model. The interdisciplinary nature of AE includes systematic research methods applicable to various engineering disciplines, e.g., computer, electrical, electronics, mechanical, and industrial engineering, to devise a common framework in order to achieve a common objective. It also involves knowledge of human and social sciences, including disciplines like sociology, psychology, marketing, cognitive, and art. These disciplines are critical to AE, and are essential for the research and analysis of people who make use of innovative technologies. Thus, applying AE in IoT-wheelchair design engineering process would result in products involving interdisciplinary domains.

C. Perception of Innovative Technologies

Technologies come and stay for a period of time in markets. One of the fundamental objectives of AE, is to understand user acceptance of innovative technologies, by applying the systematic and scientific exploration methods, that provide approximation and estimation of success rates of such innovations in markets. The interesting aspect here is to understand the different phases that constitute the process cycle of innovative technologies. This helps technical teams and financiers to formulate strategies to expand innovations effectively. Thus, for an innovation to be efficacious, the major concern for developers is to estimate the strength, weakness, and future perspective of such pioneering technologies. However, the theories, models, and systematic research methods to accomplish and implement such innovations, have not been entirely developed and explored yet.

The life cycle of innovative technology is the most important aspect for its assessment. Its thorough comprehension enables developers and investors to efficiently extend the scope of innovative technologies. Therefore, the processes and systematic methods for assessment of innovative technology should be taken into consideration. Developing IoT-wheelchair requires study of individual life cycles of both IoT and wheelchairs. after analyzing the underling architectures of both domains, we can then identify the different levels of both the system in order obtain to common interface levels for integration.

D. Early Adopter's Perception of IT

AE mainly targets information technology (IT) users, but also shifts its focus on other types of users as well. Computer technologies exist almost in every device to solve our daily tasks effectively and efficiently. For example, smart home is an intelligent/autonomous home environment setup, enables users to remotely control and manage their devices [25]. The IoT advancement has led to a wider acceptance of the concept of smart devices, and the engagement of development communities in development process is increasing rapidly, since the tools and solutions available from information communication technologies (ICTs) are scaling up every minute. Wheelchair users must have prior knowledge of sub system functions like, GPS, GSM, and other components. People with a little IT knowledge may find it difficult to operate a complex wheelchair equipped with IoT operative systems. This limits the system usage due to the lack of operating procedures. Additionally, having little knowledge may prove to be dangerous

for both the users and to the wheelchair hardware. Nonetheless, expert users of IT can operate the device with ease, and can take full advantages of the newer system functions enabled by IoT. For example, users can be guided for safe driving, informed about status of wheelchair, GPS to show the shortest routes, GSM to send any emergency requests, Wi-Fi to provide connectivity, latest updates about traffic on routes, updates about bus services, wheelchair systems information and notifications to user and remote caregiver. With the application of AE to the IoT-wheelchairs, user acceptance perception can be evaluated by recognizing and categorizing users on the basis of their IT skill levels. The IT level of wheelchair user, plays an important role while establishing an agreement between user and the wheelchair. People, who make effective use of IT to solve simple or complex problems, have been evaluated in several studies in various disciplines like ergonomics, Human Computer Interaction, and computer-based tasks with mutual interaction among participants.

Implementation of ICT in wheelchair design urges engineers again to apply acceptability engineering, to target people who actively participate to make ICT as an important part of their lives i.e., users with unique needs would require specific ICT tools and solutions. For example, some users may expect systems to offer locomotion features based on eye movement, whereas other may require gesture recognition. Additionally, others may require some form of mapping system [26] or voice controlling [27], wire or wireless monitoring that help in navigation [23]. The IoT encompasses people who are familiar with ICTs and their usage. Therefore, the steps would be prior estimation and benchmarking the IT skill levels for users, as these skills are likely to the effective utilization of IoT-wheelchairs.

E. Perception of Early Adopters' Motivations

Generally, people adopt only those technologies, which they consider as effective in improving solutions to their problems. However, different people exhibit different behavior, and pose different characteristics, so not everyone ends up in implementing the same technology for the same problem. AE aims to identify technology innovators or enthusiasts, as the key entities responsible in the innovative technology adoption process. These innovative technology enthusiasts, also called as early adopters, serve as catalysts in the technology adoption and its dissemination. The potential state-of-the-art technology adopters, whose technology acceptance decisions critically effect the adoption process, are the key sources responsible for the technology dispersion. These potential technology adopters are exclusively dependent on the technology acceptance behavior of early adopters, who direct them to either fully accept or reject newer technologies. It is difficult to predict the adoption rates of newer technologies initially, without considering the technology adoption behaviors of the early adopters. In order to fully understand and predict the recognition levels of state-of-the-art technologies, it is crucial to identify early target users, who serve as primary dispersion sources for the existing competent technologies.

In the current scenario, almost all the technological development projects, often entail in formulating some new scientific strategies and methods in the product advancement processes. This results in stimulating effects that can be directly realized in manufactured products or in the services delivered. The vital issues encountered during the planning phase, as a part of a technological implementation process, concern to users' adoption characteristics. The characteristics of Users' and technology-to-be-adopted, collectively impact the adoption behavior. Various studies have attempted to identify certain important factors responsible for assessing users' adoption behavior. Some of these factors are normative peer effect [28], self-innovativeness [29], personal characteristics, perceived technology characteristics, behavioral intentions, and managerial influence, human-related issues like time management, organizational readiness, resource limitations such as organizational users and technology key, age, full-time users vs part-time [30]. These studies contribute to measures for accessing the behavior of users towards newer technology adoption.

Typically, several studies have tried to evaluate users' technology adoption behavior by estimating three dependent variables. These are actual use, intention of use, and behavior. Engagement of user in any activity, as a result of social pressure, is termed as subjective norm. The subjective norm or social influence, which is the user's understanding of its peers' appreciation or non-recognition of the particular or universal target behavior, often succeed the above-mentioned behavioral estimation steps. Behavioral estimations, have been made in particular models only. User beliefs and social factors, are the two key elements that influence users' perception about technology acceptance and its usage. As people usually have faith in their peers or elderly people in their social network, their beliefs grow correspondingly according to the level of their peers or network connections they have. Hierarchical group dynamics and peer instigated decisions or the decisions taken by the other parties prominently impact a person's or establishments decisions in many societal and commercial situations. For example, effects could be seen in business investments, innovative technology implementations, organizations' premeditated decisions, public or private administrative voting, and custom trends.

Further, studies have proved that the social pressure by peers and elders do impact user beliefs in the technological domain. So, the users' technology acceptance rate at this juncture is dependent on the users' association with its peer or elderly network. Therefore, higher user social connections or links is directly proportionate to higher adoption rates. Observing technology adoption process through the lens of IoT-wheelchair technology, the characteristics of both the users of wheelchair and IoT-wheelchairs must be studied in relation to each other. This leads us to take up the Task-Technology Fit (TTF) model [31] as the basis for modelling the characteristics of both the users and the technology employed, which in our case are wheelchair users and IoT-wheelchairs respectively. TTF model identifies the characteristics of users and the technology, as the two important elements which ultimately lead to technology utilization. This model provides theoretical basis to know how users assess the information systems. It also tests propositions validating the background of user assessments about technologies. Theories in past studies have tried to examine the role of individuals' characteristics by using different constructs. For example, effective technology usage, experiencing herd behavior of technology stakeholders, Self-innovativeness, Mindfulness in the face of trends.

Mirmahdi [32] has identified three highly responsible factors influencing the behavior of potential adopters. These three factors are self-efficacy, personal innovativeness, and mindfulness. The ability of a person to trust himself/herself to achieve or complete a specific task is called as self-efficacy. Users with higher level self-efficacy are passionate to explore newer technologies, as they understand it as effortless service.

Self-efficacy helps in explaining users' attitude towards product purchase, online learning systems, decisioning systems, and smart devices. These sources indicate that self-efficacy acts as an analyst to forecast users' behavioral and usage intentions. Wheelchair users are the most visible groups seen in the community of disabled people. Due to varied physical impairments, their motivation levels are lower as compared to the motivation levels of common people. This indicates, that the wheelchair users exhibit lower self-efficacy, and hence are unable to achieve their objectives of using assistive technology, like wheelchair, with higher motivation. Accordingly, the IoT-wheelchair development needs designs that focus on delivering effortless services, targeted on incrementing users' self-efficacy levels, instead of engaging in the development of large complex systems. The users with higher self-efficacy levels, find technology usage easier and exhibit readiness in adopting newer technologies.

The second factor, personal innovativeness, is defined as readiness to try any information technology [33]. In Innovation Diffusion Theory, personal innovativeness of individuals determines the information sources they recognize in making decisions about technology adoption. Individuals in social systems, with higher innovative inspiration are less affected by the believes of other members, in relation to any technology adoption. Thus, potential adopters' personal innovativeness, or in other words cognitive characteristics, have a significant and positive impact on adoption of newer technologies. Personal innovativeness of wheelchair users, depend upon their literacy and intrinsic motivation levels. Wheelchair users, who show interest in exploiting newer technologies, depict higher technology acceptance in their lives. While novice users with lower literacy, would be unable to utilize and understand the complete features of the IoT-wheelchair. Therefore, the IoT-wheelchair design criteria must involve the evaluation of factors responsible for users' personal innovativeness, which could also lead to higher acceptance of risk by users' employing latest technologies.

Lastly, the third factor, Mindfulness, is a state of cognizance that consist of lively information processing, and formulation and improvement of dissimilarities, while identifying various varying perceptions. Individuals with mindful abilities are able to adapt to open environments due to their tendency towards consciousness about potential technological shifts. This encourages them to manage and handle uncertainties with lower anxiety and stress levels [33]. The stress and anxiety levels of disabled people could vary from one individual to the other. Individuals with minor disability reflect least distress patterns, whereas others with major disability problems are prone to higher cognitive load. Therefore, IoT-wheelchair users may show different behavior depending upon their cognitive load. Users with mindfulness capabilities may adopt technologies quickly, whereas other users may take longer times to adopt, due to their herd behavioral characteristics, as mentioned previously.

Observations on technology characteristics, made by the decision makers, reveal that technology characteristics significantly impact the adoption behavior of potential adopters. Technologies are accepted by users only, if it satisfies users' requirements. IoT-wheelchair designing presently requires extensive research and development to enhance current disability assistance features. Technology characteristics adequately effect on TAM relationships. In order to understand and analyze technology characteristics, this requires theoretical support to clearly explain the characteristics of technology under consideration in relation to the technology adoption. Innovation Diffusion Theory (IDT) analyzes technology characteristics in association with technology adoption process. Rogers's, IDT explains how user's observation of technology specific information stimulates its technology adoption/rejection decision. To define technology characteristics, Rogers [33] proposes five key attributes (see Table I):

TABLE I. ATTRIBUTES OF TECHNOLOGY CHARACTERIZATION [33]

Technology Improvement level	Degree of perception of a new technology, as being better than its antecedent.
Compatibility level	Degree of perception of a new technology, as being reliable with adopters' requirements.
Complexity level	Degree of perception of a new technology, as being challenging to use.
User recognition level	Degree of results of a technology, recognizable to others.
Experimentation and flexibility level	Degree of experimentation in a technology, before its adoption.

However, some researchers [34], [35] argue that only few characteristics are essential in the process of explaining and understanding the innovation technology adoption. Out of the five key attributes, the most effective ones are – relative advantage, complexity, and compatibility frequently impact adoption. Relative advantage is one of the most commonly verified characteristics and consistent analyzer of adoption behavior. A study by Moore and Benbasat [36] reveals that technological devices' relative advantage is completely related with the adoption rate. IoT and wheelchair technology relationship can thus be studied using relative advantage, which would depict the adoption behavior. Compatibility of IoT with wheelchair technology is a factor that will be a strong driver for wheelchair technology adoption, as compatibility intensely drives technology acceptance. Similarly, users' compatibility with wheelchair technology will be the other factor accountable for the increased adoption rates. Incompatibility in the dimensions of IoT-wheelchairs may lead potential technology acceptors to recognize higher insecurity in adoption and heavily shift their focus on the impacts of identity and count of potential adopters. The complexity factor in IoT-wheelchair is critical determinant of its adoption. Complex innovations demand higher learning costs for their adoption. Thus, the IoT-wheelchair complexity must be reduced to minimum levels in order to satisfy user requirements to a maximum extent. If an IoT-wheelchair system is perceived as complex by its users, then it will result in reduced attitude and motivation levels toward adoption of such newer technologies. Therefore, potential adopters' perception of technology will depend upon perceived complexity factors as is evident from the studies.

V. CONCLUSION

In this study, we proposed Acceptability Engineering framework as a standard approach to access the design perspective of engineering assistive technology, in particular wheelchair under the realm of IoT. The study highlights certain essential aspects, necessary for the design and development of wheelchairs in relation to the IoT. Using AE to extend wheelchairs with IoT, seems to be the most appropriate method for solving the issues emerging in the intercorrelation of IoT and wheelchair technology domain. The proposed model can be evaluated experimentally. Factors responsible for the increased production and market growth can be estimated. We presume industries manufacturing wheelchairs can receive potential benefits, whilst their production basis can be guided through the proposed framework. Moreover, further research is

prerequisite to enable integration of wheelchairs in the domain of IoT. Therefore, we anticipate that our proposed framework will enable research scholars, who are actively engaged in the design of assistive technologies, to widen the scope of this study from each of the highlighted perspectives.

REFERENCES

[1] R. Cruz, V. Souza, T. B. Filho, and V. Lucena, "Electric powered wheelchair command by information fusion from eye tracking and bci," in *2019 IEEE International Conference on Consumer Electronics (ICCE)*, 2019, pp. 1–2.

[2] D. P. Salgado, R. Flynn, E. L. M. Naves, and N. Murray, "The impact of jerk on quality of experience and cybersickness in an immersive wheelchair application," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[3] M. Mahmood, M. F. Rizwan, M. Sultana, M. Habib, and M. H. Imam, "Design of a low-cost hand gesture controlled automated wheelchair," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 1379–1382.

[4] A. Baltazar, M. R. Petry, M. F. Silva, and A. P. Moreira, "Driverless wheelchair for patient's on-demand transportation in hospital environment*," in *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 2020, pp. 158–163.

[5] A. Basiri, "Open area path finding to improve wheelchair navigation," *arXiv preprint arXiv:2011.03850*, 2020.

[6] T. Götzelmann and J. Kreimeier, "Towards the inclusion of wheelchair users in smart city planning through virtual reality simulation," in *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2020, pp. 1–7.

[7] J. P. Dias, B. Lima, J. P. Faria, A. Restivo, and H. S. Ferreira, "Visual self-healing modelling for reliable internet-of-things systems," in *International Conference on Computational Science*. Springer, 2020, pp. 357–370.

[8] M. A. Azad, S. Bag, F. Hao, and A. Shalaginov, "Decentralized self-enforcing trust management system for social internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2690–2703, 2020.

[9] P. P. Ray, D. Dash, and D. De, "Intelligent internet of things enabled edge system for smart healthcare," *National Academy Science Letters*, pp. 1–6, 2020.

[10] J. E. Bardram and A. Matic, "A decade of ubiquitous computing research in mental health," *IEEE Pervasive Computing*, vol. 19, no. 1, pp. 62–72, 2020.

[11] A. Haque, A. Milstein, and L. Fei-Fei, "Illuminating the dark spaces of healthcare with ambient intelligence," *Nature*, vol. 585, no. 7824, pp. 193–202, 2020.

[12] D. Z. Morgado-Ramirez, G. Barbareschi, M. Kate Donovan-Hall, M. Sobuh, N. Elayyan, B. T. Nakandi, R. Tamale Ssekitoleko, j. Olenja, G. Nyachomba Magomere, S. Daymond *et al.*, "Disability design and innovation in computing research in low resource settings," in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–7.

[13] A. Alshangiti, M. Alhudaithi, and A. Alghamdi, "Human factors in the design of wheelchair tray tables: User research in the co-design process," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 18–24.

[14] R. J. Gowran, A. Clifford, A. Gallagher, J. McKee, B. O'Regan, and E. A. McKay, "Wheelchair and seating assistive technology provision: a gateway to freedom," *Disability and Rehabilitation*, pp. 1–12, 2020.

[15] H.-C. Kim, "Acceptability engineering: the study of user acceptance of innovative technologies," *Journal of applied research and technology*, vol. 13, no. 2, pp. 230–237, 2015.

[16] S. A. Parthasarathy, S. Subash, A. Devaraj, and V. Karthik, "Design and development of arduino based multipurpose wheelchair for disabled patients," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 884–888.

[17] Q. Liu, W.-S. Zhao, and Y. Yu, "Rfid-based bidirectional wireless rollover sensor for intelligent wheelchair," *Microwave and Optical Technology Letters*, 2020.

[18] C. Wang, M. Xia, and M. Q.-H. Meng, "Stable autonomous robotic wheelchair navigation in the environment with slope way," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 10 759–10 771, 2020.

[19] S. R. Thamel, R. Munasinghe, and T. Lalitharatne, "Motion planning of novel stair-climbing wheelchair for elderly and disabled people," in *2020 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2020, pp. 590–595.

[20] H. Li, L. Bi, and J. Yi, "Sliding-mode nonlinear predictive control of brain-controlled mobile robots," *IEEE Transactions on Cybernetics*, 2020.

[21] N.-N. Dao, T.-T. Nguyen, M.-Q. Luong, T. Nguyen-Thanh, W. Na, and S. Cho, "Self-calibrated edge computation for unmodeled time-sensitive iot offloading traffic," *IEEE Access*, vol. 8, pp. 110 316–110 323, 2020.

[22] R. E. Cowan, B. J. Fregly, M. L. Boninger, L. Chan, M. M. Rodgers, and D. J. Reinkensmeyer, "Recent trends in assistive technology for mobility," *Journal of neuroengineering and rehabilitation*, vol. 9, no. 1, pp. 1–8, 2012.

[23] K. Matsuo and L. Barolli, "Prediction of rssi by scikit-learn for improving position detecting system of omnidirectional wheelchair tennis," in *International Conference on Broadband and Wireless Computing, Communication and Applications*. Springer, 2019, pp. 721–732.

[24] G. Vailland, Y. Gaffary, L. Devigne, V. Gouranton, B. Arnaldi, and M. Babel, "Vestibular feedback on a virtual reality wheelchair driving simulator: A pilot study," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 171–179.

[25] M. Ali, A. A. Ali, A.-E. Taha, I. B. Dhaou, and T. N. Gia, "Intelligent autonomous elderly patient home monitoring system," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[26] M. J. Haddad and D. A. Sanders, "Deep learning architecture to assist with steering a powered wheelchair," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2020.

[27] P. P. Dutta, A. Kumar, A. Singh, K. Saha, B. Hazarika, A. Narzary, and T. Sharma, "Design and development of voice controllable wheelchair," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2020, pp. 1004–1008.

[28] M. Conner, "Theory of planned behavior," *Handbook of Sport Psychology*, p. 3, 2020.

[29] Y. K. Dwivedi, N. P. Rana, K. Tamilmani, and R. Raman, "A meta-analysis based modified unified theory of acceptance and use of technology (meta-utaut): A review of emerging literature," *Current Opinion in Psychology*, 2020.

[30] S. Aljarboa and S. J. Miah, "Assessing the acceptance of clinical decision support tools using an integrated technology acceptance model," *arXiv preprint arXiv:2011.14315*, 2020.

[31] V. Z. Vanduhe, M. Nat, and H. F. Hasan, "Continuance intentions to use gamification for training in higher education: Integrating the technology acceptance model (tam), social motivation, and task technology fit (ttf)," *IEEE Access*, vol. 8, pp. 21 473–21 484, 2020.

[32] M. Darban and H. Amirkhiz, "Herd behavior in technology adoption: The role of adopter and adopted characteristics," in *2015 48th Hawaii International Conference on System Sciences*. IEEE, 2015, pp. 3591–3600.

[33] J. A. García-Avilés, "Diffusion of innovation," *The International Encyclopedia of Media Psychology*, pp. 1–8, 2020.

[34] L. G. Tornatzky and K. J. Klein, "Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings," *IEEE Transactions on engineering management*, no. 1, pp. 28–45, 1982.

[35] J. W. Arts, R. T. Frambach, and T. H. Bijmolt, "Generalizations on consumer innovation adoption: A meta-analysis on drivers of intention and behavior," *International Journal of Research in Marketing*, vol. 28, no. 2, pp. 134–144, 2011.

[36] G. C. Moore and I. Benbasat, "Development of an instrument to measure the perceptions of adopting an information technology innovation," *Information systems research*, vol. 2, no. 3, pp. 192–222, 1991.

Impact of Mobile Applications for a Lima University in Pandemic

Carlos Diaz-Núñez[1], Gianella Sanchez-Cochachin[2], Yordin Ricra-Chauca[3], Laberiano Andrade-Arenas[4]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades

***Abstract*—The current global pandemic situation has forced universities to opt for distance education, relying on digital tools that are currently available, such as course management platforms like Moodle, videoconferencing applications like Google Meet or Zoom, or instant messaging apps like WhatsApp. In this study it is detailed that these tools have made virtual education an effective alternative to provide education without having a physical space where teachers and students can concentrate. In addition, this document shows that in this form of teaching learning it is not necessary to have a computer, it is enough to have a cell phone to access this type of education in Peru, since most of the country's homes have a smartphone . Both students and teachers affirm that, although a little more time is invested than usual, this teaching method is satisfactory. The result obtained is that the use of mobile applications plays a very important role in virtual classes since the vast majority of students use the cell phone. In conclusion, teaching and learning in higher university education with the use of mobile applications, both teachers and students said that it was of great help due to the interaction through communication with WhatsApp, zoom, Google meet, among others. In addition, being in constant communication with the students through the applications strengthened the teaching.**

***Keywords*—Higher education; internet connection; mobile applications; pandemic; university**

I. Introduction

At the European level, m-learning is becoming increasingly common in higher education. According to the report La Sociedad de la Información en España 2016 (study prepared by Fundación Telefónica), smartphone sales in Spain reached 334.9 million in the first quarter of 2016. These figures represent 87% (data taken from the Ditrendia 2016 Report: Mobile in Spain and in the World) of the country's total number of mobile phones. In terms of connectivity, 92% of Internet users access the Internet from their smartphones [1]. In a study Wai, Ng, Chiu showed that 84.7% of university students used these applications for training purposes. They also stated that they would like to use mobile applications to search for learning materials and share information [2]. Some Latin American countries, such as Mexico and Colombia, are among the top ten countries that use these platforms [3].

With the complementary of the Information Technologies and with the easy access to Internet, it has been possible to be viable many of the things that needed a presential treatment, now it is enough to have an application in the mobile or to enter a page in Internet, to make different transactions in real time. Banks, stores, even in the area of medical care, have implemented applications for their attention from anywhere in the world.

Currently, the use of mobile devices in Peru, the majority of people who have them, which in turn presents great advantages in terms of portability. This type of e-learning is known as m-learning (mobile learning) [4]. According to a study by Futuro Labs in 2014, young people between the ages of 20 and 29 (18% of the Peruvian population), an age range in which most are university students, have mobile devices and use them on social networks [5].

In a 2017 study on m-learning student performance using the Google Classroom application, 86% of students found that it helped them improve their academic performance [6]. Another study in 2019 shows that m-learning helps university students to perform and speak in English [7].

In these times, conventional education is having gaps when it comes to teaching, since, due to factors beyond their control, the immobilization of people decreed by many governments worldwide has left them only at home, unable to go out, much less to places where many people are concentrated. Education at any level has been affected in the delivery of classes.

With the above, a virtual learning system makes e-learning the only option for providing the educational service[8]. It would also turn remote interaction into local interaction, decrease network bottlenecks and accelerate response speed [9].

E-learning and the new vision of the use of technology, provides the necessary tools of technological communications, to make the process of teaching - learning effective and viable [10].

The objective of this article is to demonstrate the impact that mobile applications had on the teaching process at a university in northern Lima in times of pandemic.

The structure of the article is formed as follows: in Section II review of the literature, explaining the background, in Section III we focus on the methodology, where we place the steps to follow, in Section IV results and discussions where the results are discussed obtained and finally Section V which is the conclusions and future work.

II. Literature Review

Currently, in times of pandemic, teaching is remotely where the use of mobile applications is relevant in education [11]. In rural areas, compared to urban areas, internet connectivity is a problem [12],In other words, what is favored in education are generally those who can access a payment plan for the use of the internet.On the other hand, the use of digital tools by students and teachers must be able to teach remote learning [13]. Students nowadays use WhatsApp more

frequently, where it is a means of communication between teachers and students, to coordinate tasks, consultations among others [14]. In addition, the use of zoom as a videoconference by both the teacher and the students, their training is important for both to know all the benefits of zoom, such as screen sharing, uploading files among others [15]. Also, the moodle platform is widely used in university education where it has its benefits for use by students and teachers. This platform allows asynchronous and synchronous interaction [16]. The Google meet is a videoconference like the zoom but each one of it has its own advantages and disadvantages, such as computer security and its elements that make it up to interact with students [17].There are other videoconferences such as Jitsi but it is not as secure as Zoom. It is also observed in students who mostly use cell phones, where they have downloaded the zoom, Google meet among others. However, there are students who use the computer that is different to the use of the cell phone. Therefore, the teacher must adapt to teach the students remotely with a mobile phone and a computer. This requires training in the use of mobile applications and digital tools [18].The LMS (learning management system) allows to have an adequate management in the teaching-learning process since you can upload files to the platform and download them in your mobile application and this is also approved by a Web Master that the administrator comes to do. of the educational entity [19] .Mobile applications in times of pandemic have allowed the use of the internet to grow exponentially due to the use of students, however those who do not benefit from the use of the mobile are the neediest students who are in poverty or extreme poverty.

III. Methodology

Next, the relevant aspects that directly influence the development of virtual education will be analyzed, such as the connectivity in the country and how many people have a smartphone- In addition, the important points of technological tools that flow in virtual education will be described. Such as Zoom, WhatsApp, Meet or Moodle Mobile applications.

A. Survey

This research will apply a survey to 15 students and 6 teachers of the University of Sciences and Humanities in the seventh semester of the career of Systems and Computer Engineering, will be conducted through a survey using Google Form tools.

The survey that will be applied to teachers and students, will use a structure in two blocks: questions on the mastery of technologies and computer resources; and questions on mobile devices in learning. The link to the forms will then be presented:

Teacher survey: https://forms.gle/hPzRAR2HgASZSUPa7

Student survey: https://forms.gle/8PYtamsVZ4U6y87H8

B. Connectivity in Peru

Internet connection is a key part of effective distance education. According to the National Institute of Statistics and Informatics (INEI), mentioned in Fig. 1, the total percentage

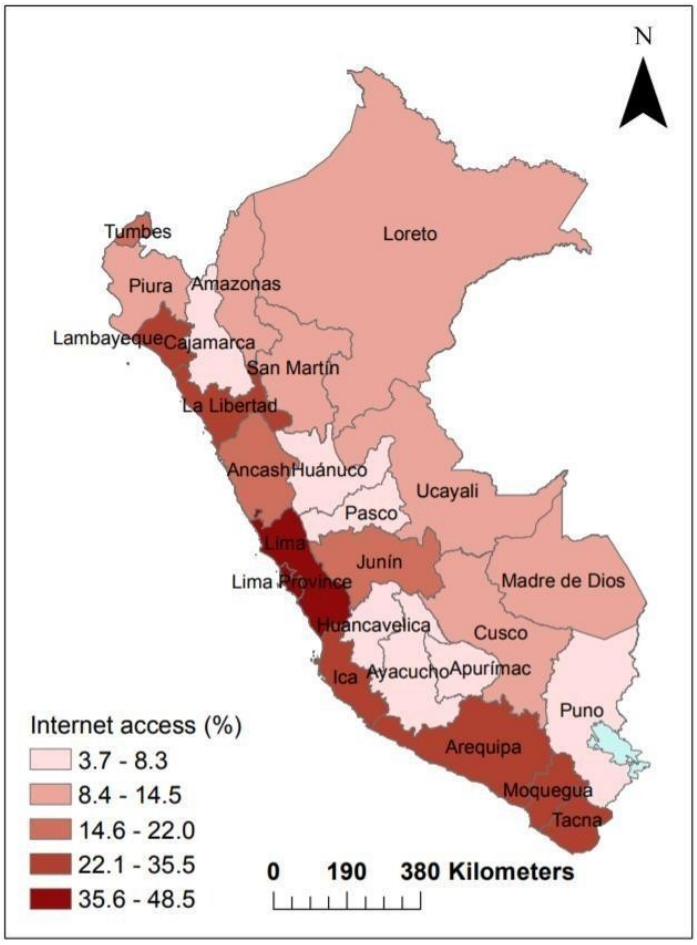


Fig. 1. Heat Map of Provinces in Peru with Internet Access.

of the population 6 years of age and older that has access to the Internet in Peru in 2016 [20].

Another important tool in the present investigation is the use of the cellular phone as the means to host the application by which the development of the courses or videoconferences would take place. As we can see in Fig. 2, the INEI gives us the following data about households in Peru that will have at least one smartphone at home by 2016 [20].

C. WhatsApp

A fast messaging experience used as an educational support tool in various aspects of connectivity. With an easy structure, basic platform for the use of this tool is practical and easy to use.

The main tools of WhatsApp and its functionality for learning are shown in Table I.

D. Moodle Mobile

Moodle Mobile is an official application of the Moodle platform, available on digital platforms, such as Windows App Store, Google Play, Market and Apple; allowing you to access it from various devices, such as a cell phone, Tablet, IPad, etc. [21].

Fig. 2. Heat Map of Provinces in Peru that have at Least one Smartphone.

TABLE I. WHATSAPP'S MAIN TOOLS

TOOL	FUNCTIONALITY
New group	Function created to join contacts in order to share useful information.%
New diffusion	You can send messages to several contacts at once about something important or event.%
WhatsApp Web	Function that allows you to use it in the browser.%
Featured Messages	Message valued with a star as it has a conversation value.%
Adjustments	Account privacy settings, chats, notifications and storage data.%

As you can see in Fig. 3, Moodle Mobile can offer the same functions as if it were a desktop, where you can all your content, courses, notes, etc.

Moodle allows us to build systems as complex as an ERP. Moreover, several activities can be carried out that will allow a very close interaction with students and teachers, such as: student and teacher collaborations, peer review and mobile learning, since it will allow them to download and configure the application [22].

Among the outstanding features that Moodle Mobile has, are:

- Contact with your course participants.
- Access their courses and download their contents; as well as monitor their own processes.
- Access links to view your course grades.
- Keeps up to date with events with the calendar.
- Feedback to teachers through surveys.
- Exams on the same cell phone.
- Learning plans for the student to check their progress.
- Teachers can grade assignments in real time.

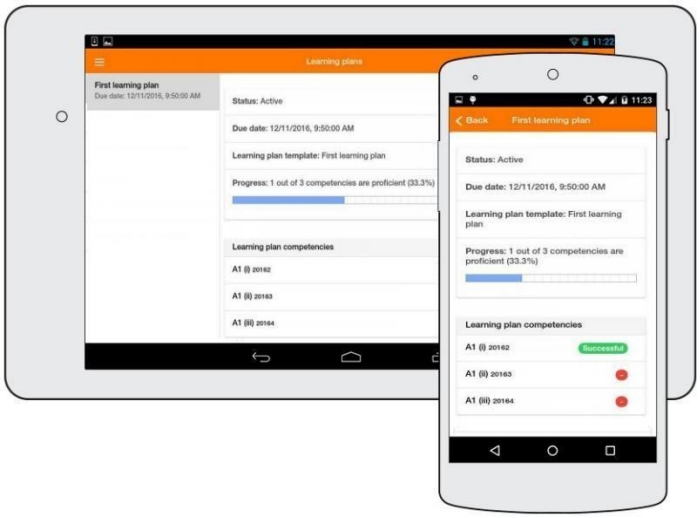


Fig. 3. Moodle Mobile Interface.

E. Zoom

Tools that allow you to work efficiently and at any time wherever you are and in an organized way to deal with the problems that arise are various collective connectivity platforms to make a virtual meeting so if you want to manage this environment is necessary to know the aspects of them and their updates that allow us to develop various expectations [23].

As shown in Table II, there are Zoom tools that are useful when conducting a class by video conference and thus be able to make the class dynamic.

TABLE II. ZOOM TOOLS

TOOL	FUNCTIONALITY
Scheduling a meeting	Allows for teleworking, videoconferencing.%
Recording a meeting	Collect data from meetings held.%
Sharing a screen	Allows several participants to do so simultaneously.%
Streaming	Broadcast live.%
Virtual blackboard	Enable a blank slate for writing or drawing.%
Live Chat	Participations through a chat.%
Management of participants	Enable and disable audio and video for participants.%
Virtual backgrounds	Allows the use of a virtual background.%

The added value of these tools is the maximum use of the experiences that allow to solve the problems of academic connectivity among others since they facilitate the generation of new knowledge creating a programmed environment with generations of new connections [22].

As shown in Fig. 4, the representation of the basic zoom tools and their easy interaction.

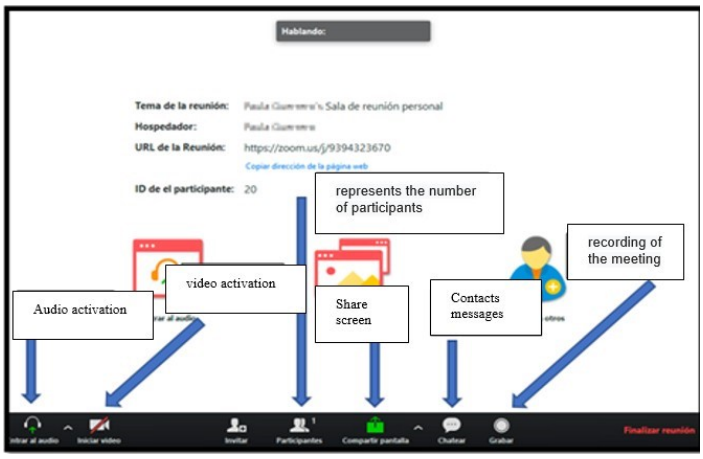


Fig. 4. Zoom Platform Tools

Besides being a free application (the basic functions), its architecture is designed to be as simple as possible, so it is not necessary to have a user's manual.

F. Google Meet

Meet is more focused on schools and businesses, is paid for and accessed from a corporate Google Suite Center account, and generates more options such as: Encrypted account, the participants can reach 100 or more depending on the plan that has been activated, helps you record meetings so you can share them and send connection links to announce the video call [24].

Google Meet is fully integrated with Google Suite and this makes it possible to join meetings directly from a Calendar event. The following Fig. 5 shows the features of the Meet interface.

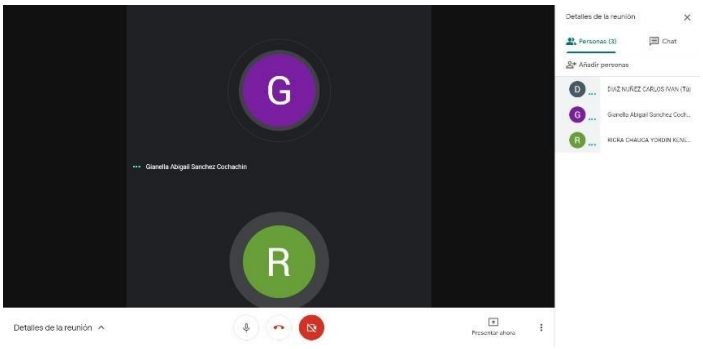


Fig. 5. Graph of Google Meet Features.

A very new feature of Google Meet is that, since May 4th 2020, Meet is available for free on the website meet.google.com and through its mobile applications [25].

Some features of Meet are that it allows screen sharing with more control and integrates screen layouts that suit users. A security feature is that the meeting host can admit or deny entry to a meeting, as well as mute or delete participants. Another is that only Google Account users join meetings, not other users.

IV. Results and Discussions

As for the teachers surveyed, the following results were obtained:

- 100% ensures that they are capable of working with mobile applications.
- Its objectives for the 2020-I semester were 100% fulfilled.
- 68% spend more than 4 extra hours than usual to prepare their classes, while 16% spend between 2 and 4 hours, and the other 16% less than 2 hours.
- 80% consider that they highly need to strengthen their digital skills, while 20% do not.
- Most teachers have mastered the Zoom, Google Meet, and WhatsApp tools, as shown in Fig. 6.
- Teachers say they always use digital tools with their students, as shown in Fig. 7.
- 66.7% of the respondents stated that the greatest difficulty during the online classes was pedagogical, while the rest had other inconveniences, as shown in Fig. 8.

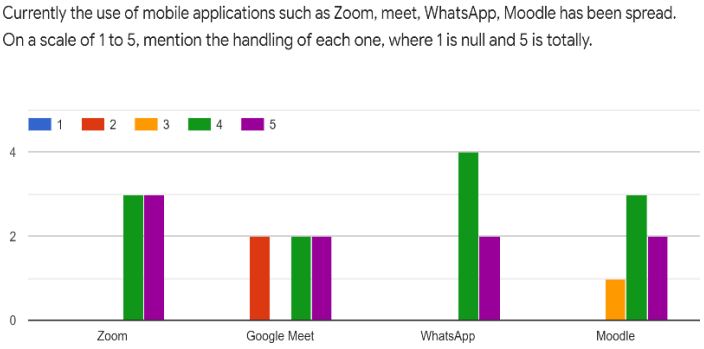


Fig. 6. Question about the use of Mobile Applications.

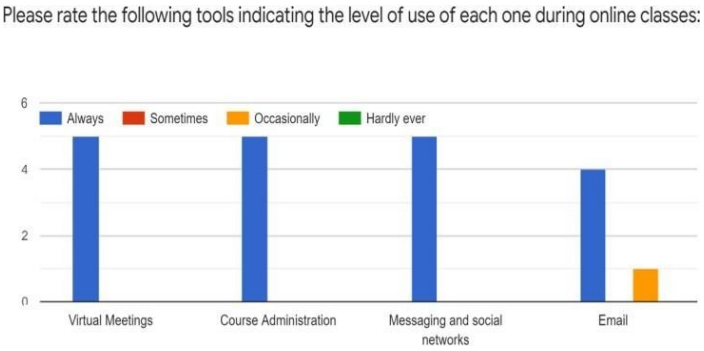


Fig. 7. Question about the use of Digital Tools.

As for the students surveyed, the following results were obtained:

- 67% consider that their teachers are trained to deal with virtual classes.

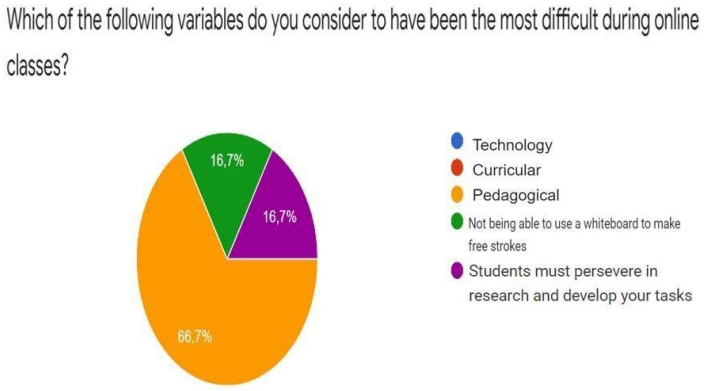


Fig. 8. Question about Difficulties in Online Classes.

- 57% say that their teachers have a high level of mastery of Moodle, Zoom, WhatsApp or Google Meet, while 47% consider it low.
- 53% are sure that mobile applications help them highly in their virtual teaching, while 40% consider it low, and only 7% interpret it as very high.
- 67% dedicate more than 4 hours of extra time than usual to their university work, while 27% dedicate between 2 and 4 hours, and the other 6% less than 2 hours.
- 93% consider that they do need to strengthen their digital skills in virtual environments, while only 7% do not.
- Most students report medium to high proficiency in the Zoom, Google Meet, and WhatsApp tools, as shown in Fig. 9.
- 87% of respondents say that mobile do motivate them for virtual learning.

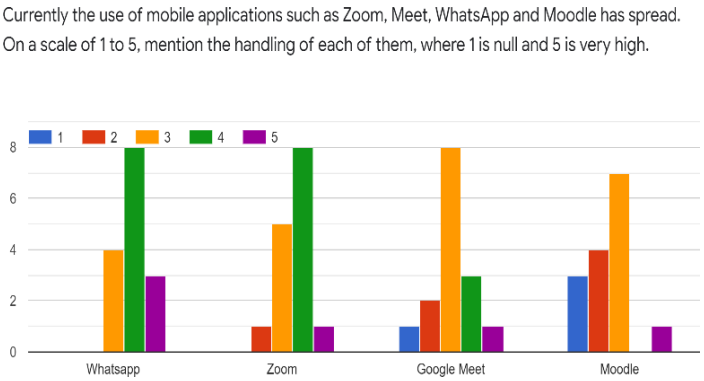


Fig. 9. Graph on the Question of use of Mobile Applications.

On the other hand, the percentages provided by INAI tell us that, although there are differences in purchasing power between one department and another, most households in Peru will have a cell phone at home by 2018, since they exceed 70%; this means being viable to transmit education in the country.

As we can see in Fig. 1, the difference in internet access is great between some departments and others. For example, while departments like Lima, Ica or the province of Callao, have a great advantage in terms of people who have access with more than 50%; others like Amazonas, Apurimac or Cajamarca, can barely reach 25% of their inhabitants. Making it clear that distance education can be given better in some regions of the country than in others.

As for the use of Moodle Mobile, it has been successful in implementation and use for the academic environment. For example, a study carried out on its use indicates that most students who used the tool did so more frequently in looking at the courses or checking the grades obtained, while the discussion forum was the least busy [16]. As can be seen in Fig. 10.

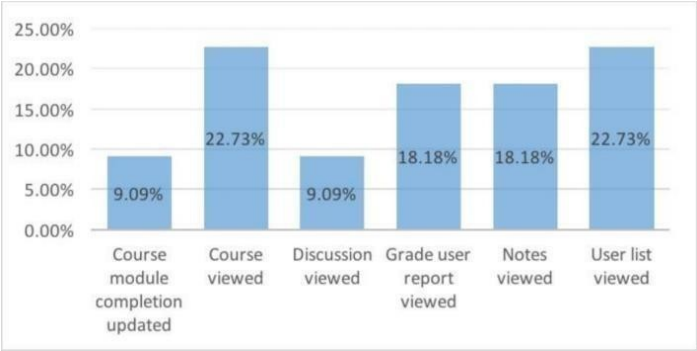


Fig. 10. Using the Moodle Mobile Tool.

In addition, the same study with a sample of 100 students, states that more than half saw it as a very easy tool to use for their courses [26], as shown in Fig. 11.

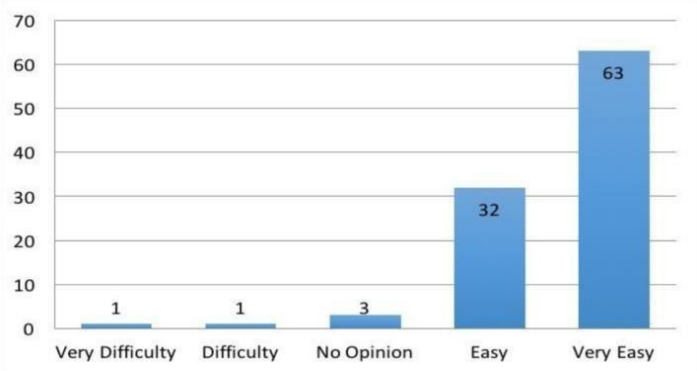


Fig. 11. Difficulty level of the Moodle Mobile Tool.

In a 2018 study, the use of WhatsApp as a means of information exchange for the academic environment was very satisfactory among university students, as they used it for academic purposes as shown in Fig. 12.

One of the most shocking questions was about the use to which it was put. As shown in Fig. 13, it mentions most frequently for records management, followed by group management.

In the XI International Congress of a University the use of the Zoom tool in the courses is very useful and beneficial for

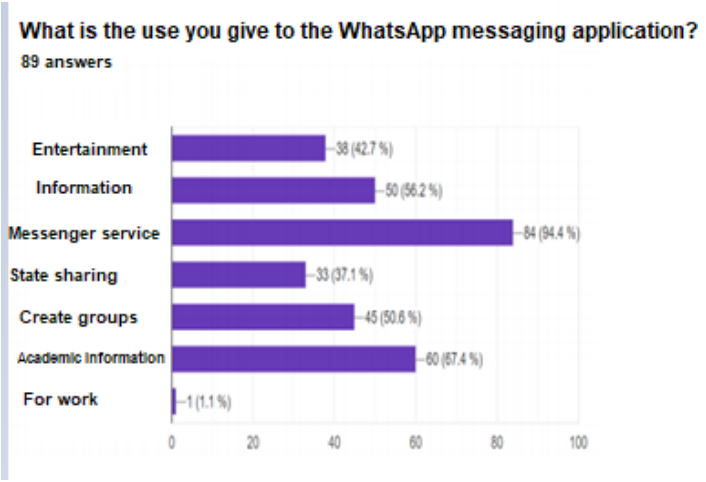


Fig. 12. . What use is Given to the WhatsApp Messaging Application?

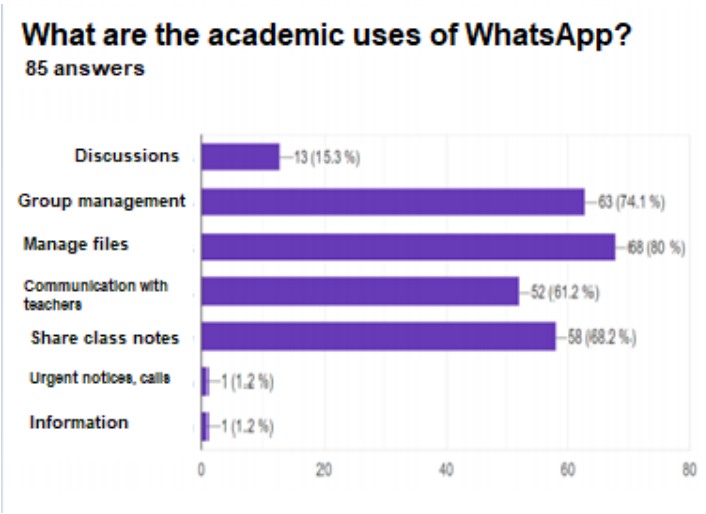


Fig. 13. Academic uses you Give WhatsApp?

the students since they are done through the videoconferences and there is enough contact with the teacher.

Another study mentions the e-learning [27] strategy of many institutions regarding the use of the Zoom and Google Meet application in education. Showing as a result as shown in Fig. 14. That the two are compatible with respect to usability, clarity and user interface.

Another 2019 study from a Latin American journal of social communication [28], mentions the good use of the WhatsApp tool in university students for academic purposes as shown in Fig. 15.

As you can relate, every year more students are using this technology tool, WhatsApp, for academic purposes and are using it for more university subjects.

In addition, Google Meet ranks among the most user friendly applications, as shown in Fig. 16.

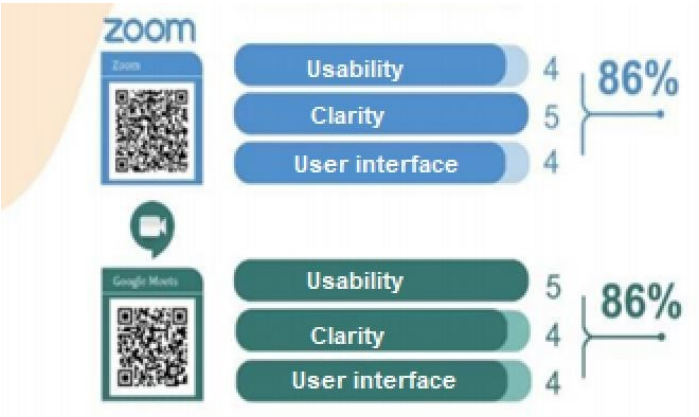


Fig. 14. Zoom and Google Meet Academic Support Strategy.

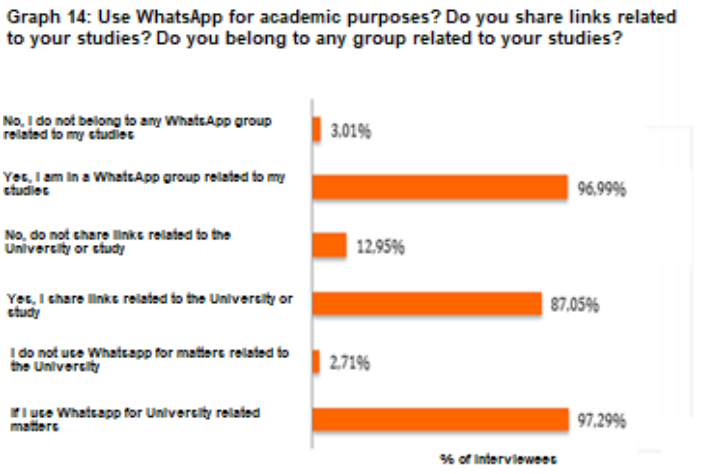


Fig. 15. Do you use WhatsApp for Academic Purposes?.

	Maximun number of participants	Can be used without installing the app	Can be used on the computer	Allows screen sharing	Ease of use
JITSI MEET	unlimited	✓	✓	✓	★★★★★
GOOGLE MEET	100	✓	✓	✓	★★★★★
FACETIME	32		✓		★★★★★
GOOGLE DUO	12	✓	✓		★★★★★
HOUSEPARTY	8	✓	✓	✓	★★★★★
DISCORD	25	✓	✓	✓	★★★★
HANGOUTS	10	✓	✓		★★★★
WHATSAPP	8				★★★★
MICROSOFT TEAMS	Between 10 and 10.000	✓	✓	✓	★★★
ZOOM CLOUD MEETINGS	100	✓	✓	✓	★★★
SNAPCHAT	16				★★★
SKYPE	10	✓	✓	✓	★★★
INSTAGRAM	6				★★★

Fig. 16. Graph Representing the Ease of use of Google Meet.

V. Conclusion and Future Work

In conclusion, regarding the results of the survey applied to both teachers and students of the University of Sciences and Humanities, m-learning is already in university and many of the teachers handle very well the mobile applications mentioned in the survey for virtual teaching. In addition, they are in constant communication with students through the applications, strengthening the teaching provided to them. Also, the expectations that were set for the semester taught were met, using the mobile applications in university teaching. On the other hand, teachers have encountered pedagogical difficulties during the development of their classes, so their improvement is a short term proposal. As for the students, they mentioned that their teachers were trained for m-learning through mobile applications, they also mentioned that using these tools help them in their learning, they are more time in constant communication with their teachers and that motivates them to continue learning even more.

The use of mobile applications in higher education has a great impact on university students, since as mentioned above in a study of the Peruvian population, there is a large percentage of the population that has a mobile device, has access to the Internet and makes use of it. In addition, many universities have conducted studies testing the use of these technological tools such as WhatsApp, Moodle Mobile, Zoom, and Google Meet in their virtual teachings and have obtained great results, academically. It is suggested as future work that a comparison is made of the impact of the use of mobile applications in university teaching in the classroom in physical, blended and remote form.

References

[1] J. C. Yáñez-Luna and M. Arias-Oliva, "M-learning: aceptación tecnológica de dispositivos móviles en la formación online," *Revista Tecnología, Ciencia y Educación*, no. 10, 2018.

[2] A. A. Vacas, J. I. N. González, and S. Á. Sánchez, "Uso de una app móvil para evaluar la calidad de la enseñanza superior," *Prisma Social: revista de investigación social*, no. 27, pp. 65–85, 2019.

[3] J. S. Mtebe and A. W. Kondoro, "Using mobile moodle to enhance moodle lms accessibility and usage at the university of dar es salaam," in *2016 IST-Africa Week Conference*, 2016, pp. 1–11.

[4] S. I. Herrera and M. C. Fénnema, "Tecnologías móviles aplicadas a la educación superior," in *XVII Congreso Argentino de Ciencias de la Computación*, 2011.

[5] F. Portilla and C. Saussure, "El uso del smartphone como herramienta para la búsqueda de información en los estudiantes de pregrado de educación de una universidad de lima metropolitana," *Educación*, vol. 25, no. 49, pp. 29–44, 2016.

[6] C. J. Fabián Coronel, "M-learning en el rendimiento académico de estudiantes de la escuela profesional de ingeniería de sistemas y computación de la universidad peruana los andes," 2019.

[7] A. García and E. Vidal, "Mobile-learning experience as support for improving the capabilities of the english area for engineering students," in *2019 International Conference on Virtual Reality and Visualization (ICVRV)*. IEEE, 2019, pp. 202–204.

[8] R. Arias-Marreros, K. Nalvarte-Dionisio, and L. Andrade-Arenas, "Design of a mobile application for the learning of people with down syndrome through interactive games," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0111187

[9] A. García and E. Vidal, "Mobile-learning experience as support for improving the capabilities of the english area for engineering students," in *2019 International Conference on Virtual Reality and Visualization (ICVRV)*. IEEE, 2019, pp. 202–204.

[10] C. Chilivumbo, "Mobile e-learning: The choice between responsive/mobile websites and mobile applications for virtual learning environments for increasing access to higher education in malawi," in *2015 IST-Africa Conference*. IEEE, 2015, pp. 1–15.

[11] A. F. Azmi, R. Nuravianty, T. I. Nastiti, and D. I. Sensuse, "Using social networking sites for learning experiences by indonesian university students," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 177–182.

[12] D. Carrillo and J. Seki, "Rural area deployment of internet of things connectivity: Lte and lorawan case study," in *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 2017, pp. 1–4.

[13] E. W. Johnson, D. Tougaw, J. D. Will, and A. Kraft, "Distance learning: teaching a course from a remote site to an on-campus classroom," in *Proceedings Frontiers in Education 35th Annual Conference*, 2005, pp. F1H–1.

[14] H. Najafi and A. Tridane, "Improving instructor-student communication using whatsapp: A pilot study," in *2015 International Conference on Developments of E-Systems Engineering (DeSE)*, 2015, pp. 171–175.

[15] J. Sutterlin, "Learning is social with zoom video conferencing in your classroom," *ELearn*, vol. 2018, no. 12, 2018.

[16] H. R. Calle and S. N. Isidro, "Use of a virtual platform as a supporting element for the acquisition of basic mathematical skills in engineering students," in *2013 8th Iberian Conference on Information Systems and Technologies (CISTI)*, 2013, pp. 1–4.

[17] J. Byrne, M. Furuyabu, J. Moore, and T. Ito, "The unexpected problem of classroom video conferencing: An analysis and solution for google hangouts and jitsi meet," *Journal of Foreign Language Education and Technology*, vol. 5, no. 2, 2020.

[18] F. M. Amin and H. Sundari, "Efl students' preferences on digital platforms during emergency remote teaching: Video conference, lms, or messenger application?" *Studies in English Language and Education*, vol. 7, no. 2, pp. 362–378, 2020.

[19] J. Y. Ardila Muñoz and E. M. Ruiz Cañadulce, "Three dimensions for learning management system (lms) evaluation," *Zona Próxima*, no. 22, pp. 69–86, 2015.

[20] C. Sotomayor-Beltran and L. Andrade-Arenas, "A spatial assessment on internet access in peru between 2007 and 2016 and its implications in education and innovation," in *2019 IEEE 1st Sustainable Cities Latin America Conference (SCLA)*. IEEE, 2019, pp. 1–4.

[21] C. Malinchi, A. Ciupe, S. Meza, and B. Orza, "A mobile exploration solution for virtual libraries in higher education," in *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, 2017, pp. 490–492.

[22] K. A. Smith, "Cooperative learning: Making "groupwork" work," *New directions for teaching and learning*, vol. 1996, no. 67, pp. 71–82, 1996.

[23] J. Wu, M. May, and C. Yang, "A moodle-based e-learning framework to conduct the manipulation skill training for an enterprise resource planning system," in *2015 IEEE 7th International Conference on Engineering Education (ICEED)*. IEEE, 2015, pp. 118–123.

[24] A. Minina and K. Mabrouk, "Transformation of university communication strategy in terms of digitalization," in *2019 Communication Strategies in Digital Society Workshop (ComSDS)*, 2019, pp. 117–120.

[25] A. Tabot and M. Hamada, "Mobile learning with google app engine," in *2014 IEEE 8th International Symposium on Embedded Multicore/Manycore SoCs*, 2014, pp. 63–67.

[26] M. M. Ujakpa, D. Heukelman, V. K. Lazarus, P. Neiss, and G. D. Rukanda, "Using whatsapp to support communication in teaching and learning," in *2018 IST-Africa Week Conference (IST-Africa)*, 2018, pp. Page 1 of 6–Page 6 of 6.

[27] C. Castro-Vargas, M. Cabana-Caceres, and L. Andrade-Arenas, "Impact of project-based learning on networking and communications competences," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110957

[28] A. Al-Omary, W. M. El-Medany, and K. J. E. Isa, "The impact of sns in higher education: A case study of using whatsapp in the university of bahrain," in *2015 Fifth International Conference on e-Learning (econf)*, 2015, pp. 296–300.

Deep Convolutional Neural Network for Chicken Diseases Detection

Hope Mbelwa[1], Dina Machuve[3]
School of Computational and
Communication Science and Engineering
The Nelson Mandela Institution of Science and Technology
Arusha, Tanzania

Jimmy Mbelwa[2]
College of Information and Communication Technologies
University of Dar es salaam
Dar-es-salaam, Tanzania

***Abstract*—For many years in the society, farmers rely on experts to diagnose and detect chicken diseases. As a result, farmers lose many domesticated birds due to late diagnoses or lack of reliable experts. With the available tools from artificial intelligence and machine learning based on computer vision and image analysis, the most common diseases affecting chicken can be identified easily from the images of chicken droppings. In this study, we propose a deep learning solution based on Convolution Neural Networks (CNN) to predict whether the faeces of chicken belong to either of the three classes. We also leverage the use of pre-trained models and develop a solution for the same problem. Based on the comparison, we show that the model developed from the XceptionNet outperforms other models for all metrics used. The experimental results show the apparent gain of transfer learning (validation accuracy of 94% using pretraining over its contender 93.67% developed CNN from fully training on the same dataset). In general, the developed fully trained CNN comes second when compared with the other model. The results show that pre-trained XceptionNet method has overall performance and highest prediction accuracy, and can be suitable for chicken disease detection application.**

***Keywords*—Image classification; Convolutional Neural Networks (CNNs); disease detection; transfer learning**

I. Introduction

The poultry sector in Tanzania is economically significant, supporting up to 37 million households. The farmers in the country keep different birds, whereby chicken account for 96% of all livestock in the country [1], [2]. However, the growth rate of the poultry population is low, with an average of 2.6% annual growth in Tanzania mainland [1]. Production is greatly affected by different challenges like unreliable markets, scarce inputs [3], [4], shortage of timely extension information [5] and devastating diseases like Newcastle, Coccidiosis and Salmonella [6].

Coccidiosis is caused by parasites of the genus Eimeria that affects the intestinal tracts of poultry. Chicken are host to seven species of Eimeria. Coccidiosis is ranked as a leading cause of death in poultry with Eimeria tenella (E.tenella) among the most pathogenic parasite [7]. The typical diagnostic procedure involves counting the number of oocysts (expressed as oocysts per gram [opg]) in the droppings or examining the intestinal tract to determine the lesion scores [8].

Salmonella are bacterial pathogens of genus Salmonella that causes the disease to poultry and humans. Salmonella pullorum (SP) and Salmonella gallinarum (SG) pathogens cause pullorum disease and fowl typhoid in poultry, respectively. Salmonella enteritidis (SE) and Salmonella typhimurium (ST) strains are associated with human infections transmitted through the food-chain of poultry and poultry products [6]. Polymerase Chain Reaction (PCR) procedure is used for the detection and identification of the various Salmonella strains. The diseases have a significant negative economic impact on poultry farmers resulting to high economic losses.

Disease diagnostics in chicken involves different methods including counting the number of oocytes in the stool or intestinal scrapings [7], [8], isolation and identification and PCR procedures which takes several diseases to diagnose. The main way of transmission of the diseases is through contaminated feed, excretions from infected chicken or oral via the navel. The clinical signs in the infected chicken may be either digestive or respiratory. This work focuses on the digestive signs of the diseases because occurrence of the disease in the chicken influences the colour of the droppings. The digestive clinical sign of chicken infected with coccidiosis is severe blood/ brown diarrhea; salmonella is white diarrhea.

Images are artefacts that depict visual perception. They have been used for diagnosis and detection in various fields including medical, agricultural and other fields [9], [10]. There are different existing image datasets that are on the cloud accessed when training different models. These datasets include Fashion-MNIST, CIFAR-10, ImageNet to mention a few. Various computer vision studies have used image datasets in either classification, detection, recognition and segmentation of different research problems. Recently, the character portrayal is commonly used in disease diagnostics using images. Different levels of features are captured and analyzed based on various aspects, including colour. Normally, classification quality depends on feature presentations in the images. Images are preprocessed (labelled and tuned) in order to maintain the realization of the former facts. Therefore, a data-driven approach for image classification is more robust towards the variety of image attributes and diseases. Albarqouni et al. [9] used breast cancer histology images for detection of breast cancer disease; also Zhang et al. [11] applied ultrasound images in the detection of ovarian tumours. Ashraf et al. [12] worked to improve disease diagnostics using different body parts image dataset. Similarly, In the agricultural field, researchers in [10], [13]–[15] have created and used leaf image datasets in the diagnosis of diseases in different plants like tomato, cassava, bananas and wheat.

Classical machine learning techniques have been used in earlier studies for disease detection and classification. Sadeghi et al. [16] used Support Vector Machine (SVM) and Decision Tree to detect sick chickens infected with clostridium perfringens using the sound they make. In their study, vocals from both health and unhealthy chickens were taped. Facets were extracted and used to train the classifiers, whereby the accuracy of the neural network increased gradually up to 100%. Zhuang et al. [17] also used SVM approach to detect sick broilers infected with bird flu. Their work proposed an algorithm to classify the isolated inoculated broilers based on the analyzed structures and features, and the algorithm attained an accuracy rate of 99% when evaluated on the test data. Hepworth et al. [18] predicted the regularity occurrence of hock burn, swelling skin around the hock in broiler chicken. Data from farms were collected in a period of over 36 months and, learned variables of dependency were extracted, and a classifier was trained to attain an accuracy of 78%. In work by Hemalatha et al. [19] SVM was used to diagnose avian pox in chicken, images of chicken from the farm were collected then split into training and test sets. The classifier was trained on the data and obtained an accuracy of 92.7%.

Despite good prediction performance of the classical machine learning approaches, Ferentinos et al. [20] presented that traditional machine learning techniques are constrained in images and features processing. The deep learning techniques have gained more attention in computer vision and image classification, particularly in enhancing the performance of image classification and retrieval as opposed to traditional machine learning approaches. Therefore, in this research a Convolution Neural Network (CNN) is used due to the following reasons: (i) It involves multi-layer processing. (ii) It allows optimization of the extracted features. (iii) It is fast and requires less computational power. Deep Convolution Neural Networks (DCNN) enable the computer to interpret captured data objects (feature extraction and representation) for classification, localization and recognition to be automatically learned [21]. It has been used for early disease detection in both plants (crops) and animals [10].

Transfer learning (TL) refers to the use of a known model used in a previously known dataset to another application in the same machine learning domain. For computer vision, transfer learning is widely used in different applications. The most known and common pre-trained models include VGG, Resnet, Inceptionnet and other well-known models [22]. The idea of using pre-trained models makes a considerable revolution in the field of Artificial Intelligence enabling models to be developed from very little data. There are two main advantages when using TL [23]. First, it performs well on both large and smaller datasets. Secondly, it is easy to reduce overfitting of the model with larger dataset when it is applied to pre-trained model [23]. Since our problem is a computer vision problem, we use TL, and also we develop our CNN architecture, and compare the results with the best pre-trained model.

The objective of this research is to build a model for early detection and classification of diseases in chicken using our developed dataset of faecal images collected in different poultry farms and inoculation sites.

The rest of the paper is organized as follows: Section II describes the materials and methods used for the disease detection, Section III gives the results and discussion lastly, Section IV concludes the study and presents future work.

TABLE I. DATASET SPLITTING TRAINING, VALIDATION AND TESTING SETS.

Class	Image in each Class	Training	Validation	Test
Health	508	305	102	102
Coccidiosis	516	310	103	103
Salmonella	566	340	112	113
Total Images	1590	955	317	318

II. MATERIALS AND METHODS

A. Experimental Setup

The model is generated from the Kaggle Environment, and we run the training set under TPU-v3.8 environment using python v3.7 in a 16 GB RAM computer. We avoid to alter the images since we have collected data using mobile phone cameras and assumption made is the end-user will use unaltered images for prediction. However, we stated earlier that the images need to be converted to either 224×224 or 512×512 pixels as shown in the Table II.

B. Dataset

In this work, we collected faecal images from small scale farmers and inoculation sites in Kilimanjaro and Arusha regions from February to June 2020. We used different mobile phone cameras with different resolutions and images were in a Joint Photographic Group (JPG) format. A total of 1590 images were collected and distributed to 3 class labels. Health 508 images, Coccidiosis 516 images and Salmonella 566 images. We managed to have a dataset that meets our exact specifications of different images, as shown in Fig. 1. The images were then split into training 60%, validation 20% and testing 204% that is 955 images for training, 317 images for validation and 318 images for testing as illustrated in Table I.

C. Pre-Processsing

Deep Learning is a class of machine learning algorithms inspired by the structure of the human brain. Deep Learning algorithms use complex, multilayered neural networks, where the level of abstraction increases gradually by non-linear transformations of input data. DL demonstrates a high ability to solve image classification problems since it learns from the image features. Image classification techniques are naturally based on two stages: one is the mining of features, and the other is the classification component.

Images in the dataset are labelled as per respective classes then converted into tensor records [24] in different resolutions 224×224 and 512×512 pixels and compressed in order to minimize training time. The advantage of resizing the images is to achieve reasonable resolution since a constant input dimensionality of the data is needed by CNN to train the optimized model [25].

(a) (b) (c)

Fig. 1. Sample Images from the Fecal Image Dataset. (a) Coccidiois. (b) Health. (c) Salmonella.

D. Proposed Model

This study uses architectures designed for image classification including VGG [26], Resnet [27], XceptionNet [28] and Mobile net [29]. The VGG 16 and 19 refer to the number of weighted layers in each network, consisting of blocks with incremental convolution layers with a 3×3 size filters, a high number of parameters and requires a considerable amount of time to train [26]. A study by [30] proposed a TL strategy for image recognition using the VGG 16 architecture, where it was trained on a small dataset and achieved an accuracy of about 87%. They compared the results with training the same dataset on a fully untrained VGG 16 architecture and attained an accuracy of 46%. This supports the proposed argument on using a pre-trained CNN in our work. In addition, [31] used a pre-trained VGG 16 and 19 architectures on palm vein recognition problem with an accuracy of 90% and 92% respectively. Resnet architecture is built up with residual connections that are networks within networks. Different Resnets are available varying in numbers of layers, commonly used is the Resnet 50 that consists of 50 layers. Despite the fact that it has more layers than the VGG, it consumes less memory when training [27]. XceptionNet architecture optimizes the convolutions in inception so that they consume less memory during training [28]. MobileNet is an efficient architecture for models deployed in mobile devices. It consists of trivial separable layers of neural networks created from depth-wise distinguishable convolution filters. The mechanism behind each input network is from one convolution filter that is 1×1 convolutions [32]. A study conducted by [33] compared pre-trained models including Resnet 50, Mobilenet and VGG 19 for diagnosis of Pneumonia disease. The architectures had an accuracy of 87%, 92% and 90% respectively. From the previous studies, it is evident that no architecture performs well on every problem; this motivates us to apply TL on pre-trained models to solve our problem as well as develop model based on CNN architecture from scratch.

In this study, we propose two solutions which are based on convolutional neural network. In the first method we design the CNN from scratch and then we train the proposed CNN to obtain the model for chicken disease detection. Secondly, we leverage the use of transfer learning by optimizing through pre-training and fine-tuning the pre-trained models.

1) CNN architecture: Our proposed CNN architecture involves stacking of multi convolution layers, and the whole architecture is given in Fig. 2. In the first layer, the image with size either 224×224 RGB or 512×512 is fed to the stack of convolution layer as input. The convolutional layers have filters with the small receptive fields of 3×3 and are followed by max-pooling layer, which performs over a 2×2 pixel window. These layers form a single block, and we repeatedly apply the block by increasing the depth of filters in the network in such as $32, 64, 64, 128, 128, 256, 256, 512$ for the full convolution blocks. In each block, the same padding is applied to maintain the height and the width shape of the output features maps matching the inputs features. ReLU activation is used for all layers; meanwhile,he_uniform is considered for weight initialization for all blocks. During training, normal stochastic gradient descent is used to minimize the error, and in evaluation, we leverage log loss and accuracy as the metrics. We have noticed that some of the images from the dataset may contain more than one disease; hence categorical cross-entropy loss function seems to fit our problem with the log loss as an evaluation metric. An output layer with three nodes and softmax activation is used since the problem is multi-class classification. Softmax is the right choice because the output from the node is the likelihood for the output to be either of the three classes. Schedule learning rate is used in the experiment with some callback features.

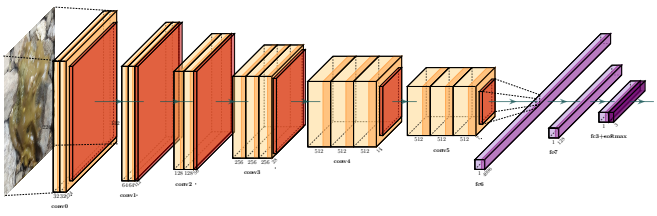


Fig. 2. Configuration of Fully Connected CNN Proposed Model.

2) XceptionNet: For this study, we use XceptionNet, an interpretation of Inception modules in convolution neural networks, introduced in [28]. Our proposed method is the use of pre-trained models that were originally trained on very large datasets including, ImageNet. Using pre-trained models is an added advantage to researchers as it uses the minimum time during training as compared to fully trained models that are built from scratch. We select the XceptionNet architecture based on our problem because it gives the best results as compared to others, we also consider the weights,

TABLE II. THE SIZES OF IMAGES IN THE DATASET CHANGED TO 224×224 AND 512×512

Resolution	Coccidiosis	Health	Salmonella
224×224			
512×512			

time consumed during training and accuracy it achieves [34].

The XceptionNet architecture performs depth-wise separable convolutional operations, whereby each channel has only one kernel to perform convolution; hence computational complexity is minimized. Each kernel is a 2-dimensional and is convoluted over a single channel. From the architecture, as a lightweight network, XceptionNet not only involves lesser number of residual blocks than other CNN models but also has a stronger classification impact over other CNN models because the number of parameters and weights is less. Most image recognition models have broad parameters and a large number of calculations that are not ideal for embedding in mobile devices. For the identification of chicken diseases, we must also consider how to quickly and accurately identify coccidiosis and salmonella in areas where there is limited access to robust tools for disease diagnosis. This is why we suggest XceptionNet architecture in our study.

III. RESULTS AND DISCUSSION

A. Experimental Data

The digestive signs of the diseases are significant for our study based on the data (chicken droppings images). We gathered the data in the field with the help of a veterinary officer to correctly identify the features in the data we need. We collected a dataset of 1590 images for both healthy and infected images, as shown in Table III. To make data more inclusive, we randomly generated the training and test sets as presented in Table I.

TABLE III. COLLECTED DATASET

Class	Image in each Class
Health	508
Coccidiosis	516
Salmonella	566
Total Images	1590

TABLE IV. HYPERPARAMETERS USED FOR TRAINING FULLY CNN MODEL

Parameter	Value
Learning rate maximum	0.000012
Learning rate minimum	0.00001
Maximum learning rate attain at epoch	200
Learning rate exponential decay rate after 10 epochs	0.6

B. Classification Results and Analysis

When training the model, we augment the images by using different techniques in order to increase the size of the dataset. The significant number of dataset has two potential benefits: first, to avoid over-fitting and secondly to make the model learn from unseen datasets. The augmentation techniques used in our model are image flipping, image cropping and padding, and add random image saturation.

For the proposed model based on CNN architecture without using pre-trained model, we develop the model using the parameter specified in Table IV.

The result shows the CNN perform better compared to

TABLE V. PERFORMANCE COMPARISON FOR DIFFERENT ALGORITHMS

Method & Algorithm	Log loss	Validation Accuracy	Validation Loss
VGG 16	0.35	0.8933	0.3522
Resnet 50	4.8	0.3133	1.1131
MobileNet	0.89	0.6833	0.9005
XceptionNet	0.15	0.94	0.161
CNN	0.20	0.9367	0.2282

TABLE VI. HYPERPARAMETERS USED FOR TRAINING XCEPTIONNET MODEL

Parameter	Value
Learning rate maximum	0.000012
Learning rate minimum	0.00001
Maximum learning rate attain at epoch	10
Learning rate exponential decay rate after 10 epochs	0.8

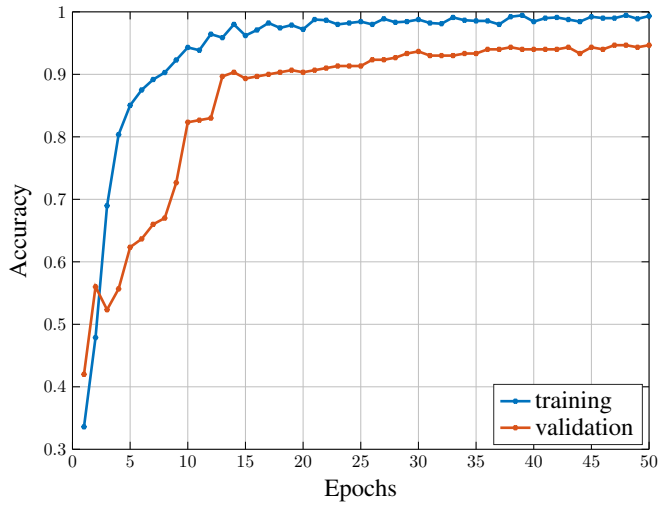
other the pre-trained models used, as shown in the Table V except it under-performs when compared to XceptionNet that will be discussed in the next paragraph. From the Table V it can be clearly seen that for all performance metrics considered, the proposed model based on CNN perform outperform other models, however lower than XceptionNet. For the case of XceptionNet, the results are presented in the Table V as well as other results.

Basically, we use the XceptionNet with modification which is presented as follows; First, during training and class prediction, the first layer of XceptionNet is removed, and we set the input layer based on the image size pixel from our dataset. Also, the last layer is flattened, and Pool 2D is applied, followed by a dense layer with the softmax activation with 3 output nodes. The summary of parameters and hyperparameters used when training the pre-trained model is given in the Table VI.

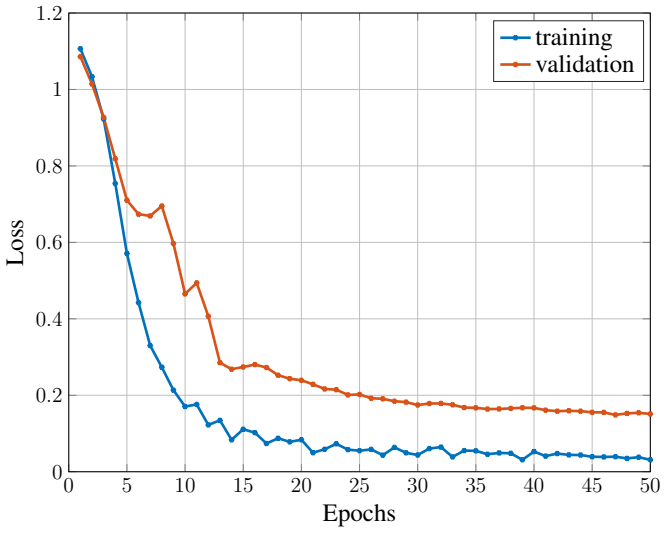
This section presents the results obtained after evaluating the performance of the pre-trained models. Table I briefly shows the summary of the dataset and how the data is split into the training set, validation set and the test set. We start by showing the performance of other pre-trained models used in our work, and then our proposed CNN architecture model, as shown in Table V. Since we use TPU-based framework, it only takes 3.95min for training our dataset. We run 50 epochs, and each epoch has 8 steps per epoch, which results in 50×8 iterations to complete the training process. In the Fig. 3(a) and Fig. 4(a), it can be seen that the accuracy of validation keeps increasing and converges in the 50^{th} epoch. The results show that the XceptionNet with the hyperparameters shown in the Table VI outperform other models using 0.04seconds for a single prediction .

IV. CONCLUSION

In this study, we present the novel chicken disease detection method using Transfer Learning approach on a pre-trained CNN. The key elements that can improve the performance of the extension officers and poultry farmers in the early detection of chicken diseases are the use of computer-aided instruments and accurate data. The creation of such image processing techniques that can assist farmers is a necessity of the present



(a) Accuracy

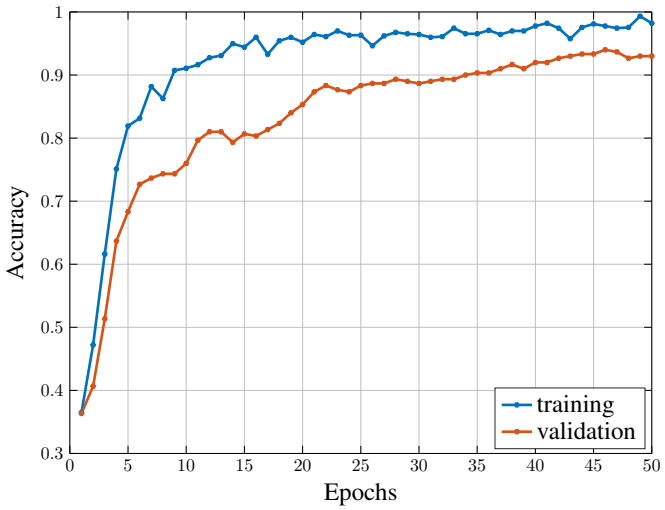


(b) Loss

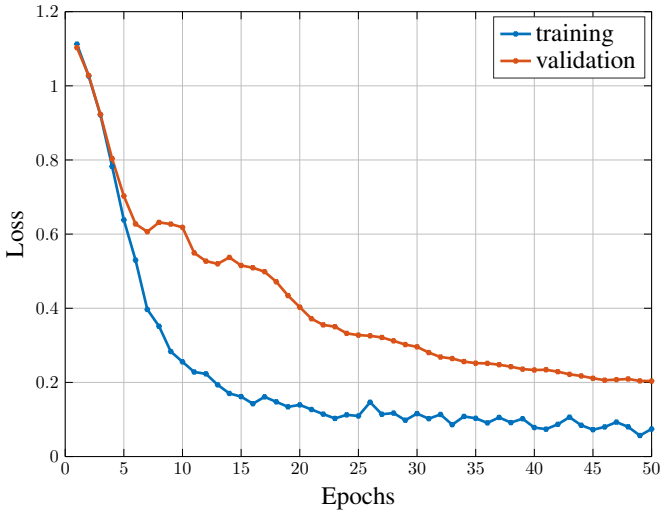
Fig. 3. The Training and Validation set (a) Accuracy and (b) Loss of the XceptionNet Evaluated on Dataset with 512×512 Resolution.

era. These methods are valuable to reduce the losses incurred and increase productivity, and it is clear that the diseases can be detected at an early stage before they lead to deaths of the chicken. Computer vision work has been trying to reduce the gap for the past few decades by designing automated systems that can process images for decision making using computers. Specifically, we generate the CNN model, which learns the hidden pattern among the different faecal images in our dataset. The supervised learning algorithm predicts the three categories which we named them as coccidiosis, health and salmonella. The results obtained show the proposed model achieved up to 94% accuracy. In comparison to the VGG, Resnet and Mobile net architectures, our proposed solution outperforms them by far. The experimental findings indicate that our method works well on diseases detection in the chicken and can be used for robust diagnostics.

In the future, we aim to collect more faecal images to



(a) Accuracy



(b) Loss

Fig. 4. The Training and Validation Set (a) Accuracy and (b) Loss of the XceptionNet Evaluated on Dataset with 224×224 Resolution.

expand the dataset hence room for future studies on other chicken diseases using the dataset. The developed model will be deployed in mobile devices for easy interaction by the end-users.

ACKNOWLEDGMENT

The authors would like to thank the African development Bank (AfDB) Project ID No: P-Z1-IA0-016 for funding this work.

REFERENCES

[1] FAO, "Poultry development," 2013. [Online]. Available: http://www.fao.org/3/i3531e/i3531e.pdf

[2] M. of Livestock and Fisheries, "Tanzania livestock modernization initiative," 2015. [Online]. Available: MinistryofLivestockandFisheries.(2015).TanzaniaLivestockModernizationInitiative.MinistryofLivestockandFisheriesDevelopment:UnitedRepublicofTanzania.

[3] D. Damena, M. Kidane, R. Alemu, M. Sombo, A. Fusaro, A. Heidari, T. Chibssa, and H. Chaka, "Characterization of newcastle disease virus and poultry-handling practices in live poultry markets, ethiopia," *SpringerPlus*, vol. 3, p. 459, 08 2014.

[4] J. Wong, J. de Bruyn, B. Bagnol, H. Grieve, M. Li, R. Pym, and R. Alders, "Small-scale poultry and food security in resource-poor settings: A review," *Global Food Security*, 05 2017.

[5] E. Lwoga, P. Ngulube, and C. Stilwell, "Information needs and information-seeking behaviour of small-scale farmers in tanzania," *Innovation: Journal of Appropriate Librarianship and Information Work In Southern Africa*, vol. 40, pp. 82–103, 09 2010.

[6] T. Desin, W. Köster, and A. Potter, "Salmonella vaccines: past, present and future," *Expert Review of Vaccines*, vol. 12, pp. 87–96, 01 2013.

[7] L.-S. Lim, Y.-L. Tay, H. Alias, K.-L. Wan, and P. H. Dear, "Insights into the genome structure and copy-number variation of eimeria tenella," *BMC genomics*, vol. 13, p. 389, 08 2012.

[8] G. Grilli, F. Borgonovo, E. Tullo, I. Fontana, M. Guarino, and V. Ferrante, "A pilot study to detect coccidiosis in poultry farms at early stage from air analysis," *Biosystems Engineering*, 02 2018.

[9] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1313–1321, May 2016.

[10] G. Owomugisha, J. Quinn, E. Mwebaze, and J. Lwasa, "Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease," 06 2019.

[11] Z. Zhang and Y. Han, "Detection of ovarian tumors in obstetric ultrasound imaging using logistic regression classifier with an advanced machine learning approach," *IEEE Access*, vol. 8, pp. 44 999–45 008, 2020.

[12] R. Ashraf, M. Habib, M. Akram, M. Latif, M. Malik, M. Awais, S. Dar, T. Mahmood, M. Yasir, and Z. Abbas, "Deep convolution neural network for big data medical image classification," *IEEE Access*, vol. PP, pp. 1–1, 06 2020.

[13] A. El-Kereamy, J. Kreuze, Z. Yin, D. Hughes, A. Ramcharan, K. Baranowski, P. Mccloskey, B. Ahmed, and J. Legg, "Deep learning for image-based cassava disease detection," *Frontiers in Plant Science*, p. 1852, 10 2017.

[14] G. L. Grinblat, L. C. Uzal, M. G. Larese, and P. M. Granitto, "Deep learning for plant identification using vein morphological patterns," *Computers and Electronics in Agriculture*, vol. 127, pp. 418 – 424, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169916304665

[15] J. Ubbens, M. Cieslak, P. Prusinkiewicz, and I. Stavness, "The use of plant models in deep learning: an application to leaf counting in rosette plants," *Plant methods*, vol. 14, no. 1, p. 6, 2018.

[16] M. Sadeghi, A. Banakar, M. Khazaee, and M. Soleimani, "An intelligent procedure for the detection and classification of chickens infected by clostridium perfringens based on their vocalization," *Revista Brasileira de Ciência Avícola*, vol. 17, pp. 537–544, 10 2015.

[17] X. Zhuang, M. Bi, J. Guo, S. Wu, and T. Zhang, "Development of an early warning algorithm to detect sick broilers," *Computers and Electronics in Agriculture*, vol. 144, pp. 102–113, 01 2018.

[18] P. Hepworth, A. Nefedov, I. Muchnik, and K. Morgan, "Broiler chickens can benefit from machine learning: Support vector machine analysis of observational epidemiological data," *Journal of the Royal Society, Interface / the Royal Society*, vol. 9, pp. 1934–42, 02 2012.

[19] Hemalatha, S. Muruganand, and R. Maheswaran, "Recognition of poultry disease in real time using extreme learning machine," 2014.

[20] K. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 02 2018.

[21] S. Kumar, A. Pandey, S. Kondamudi, S. Kumar, S. Singh, A. Singh, and A. Mohan, "Deep learning framework for recognition of cattle using muzzle point image pattern," *Measurement*, vol. 116, pp. 1–17, 10 2017.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] H. Lee, S. Eum, and H. Kwon, "Is pretraining necessary for hyperspectral image classification?" in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 3321–3324.

[24] A. Ramcharan, P. McCloskey, K. Baranowski, N. Mbilinyi, L. Mrisho, M. Ndalahwa, J. Legg, and D. Hughes, "Assessing a mobile-based deep learning model for plant disease surveillance," *CoRR*, vol. abs/1805.08692, 2018. [Online]. Available: http://arxiv.org/abs/1805.08692

[25] W. Wang, Y. Hu, T. Zou, H. Liu, J. Wang, and X. Wang, "A new image classification approach via improved mobilenet models with local receptive field expansion in shallow layers," *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1–10, 08 2020.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2014.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017.

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[30] C. Iorga and V. Neagoe, "A deep cnn approach with transfer learning for image recognition," in *2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2019, pp. 1–6.

[31] M. Wulandari, Basari, and D. Gunawan, "Evaluation of wavelet transform preprocessing with deep learning aimed at palm vein recognition application," *AIP Conference Proceedings*, vol. 2193, no. 1, p. 050005, 2019. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.5139378

[32] W. Wang, Y. Li, T. Zou, X. Wang, J. You, and Y. Luo, "A novel image classification approach via dense-mobilenet models," *Mobile Information Systems*, vol. 2020, pp. 1–8, 01 2020.

[33] Z. Yue, L. Ma, and R. Zhang, "Comparison and validation of deep learning models for the diagnosis of pneumonia," *Computational Intelligence and Neuroscience*, vol. 2020, 2020.

[34] M. T. Almalchy, S. Monadel Sabree ALGayar, and N. Popescu, "Atrial fibrillation automatic diagnosis based on ecg signal using pretrained deep convolution neural network and svm multiclass model," in *2020 13th International Conference on Communications (COMM)*, 2020, pp. 197–202.

Efficient Lung Nodule Classification Method using Convolutional Neural Network and Discrete Cosine Transform

Abdelhamid EL HASSANI[1], Brahim AIT SKOURT[2], Aicha MAJDA[3]
Sidi Mohamed Ben Abdellah University - Faculty of Science and Technology, BP 2202 Fez Morocco

***Abstract*—In today's medicine, Computer-Aided Diagnosis Systems (CAD) are very used to improve the screening test accuracy of pulmonary nodules. Processing, classification, and detection techniques form the basis of CAD architecture. In this work, we focus on the classification step in a CAD system where we use Discrete Cosine Transform (DCT) along with Convolutional Neural Network (CNN) to perform an efficient classification method for pulmonary nodules. Combining both DCT and CNN, the proposed method provides high-level accuracy that outperforms the conventional CNN model.**

***Keywords*—Convolutional neural network; discrete cosine transform; pulmonary nodule classification; computer aided diagnosis systems**

I. Introduction

Cancer incidence and mortality are increasing rapidly around the world [1]. It represents the second leading cause of death globally, and it was responsible for an estimated 9.6 million deaths in 2018. Lung cancer is the most frequently diagnosed cancer in both genders, and it is the leading cause of cancer-related death worldwide, with over 2.09 million cases. Even worse, Lung cancer is killing over 1.76 million people yearly (according to the World Health Organization), which represents 20% of the overall cancer-related deaths [2].

According to the American Cancer Society, most cases of lung cancer are diagnosed at a late stage when it is already metastasized as symptoms usually appear until a late stage. Early detection of suspected pulmonary nodules is very important and it could potentially increase survival rates. There are several types of medical imaging modalities for lung cancer screening, but the most frequently used for nodule detection and analysis is Computed Tomography (CT) [1].

In many cases, it is difficult to obtain an accurate diagnosis due to the complicated morphological structure of nodules. A pulmonary nodule is simply an oval-shaped spot growth in the lung. Its form can be confused with other shapes in a CT-scan like end-on vessels. A Nodule is called pulmonary mass when its diameter is larger than 3 centimeters, otherwise it's called pulmonary nodule. It is also called micronodules when the diameter is smaller than 4 millimeters. Countless amount of nodules can be discovered during a screening test, and each one of them can be either malignant (cancerous) or benign (noncancerous). The figure 1 shows an example of nodules on two different CT-scans from LIDC database.

To deal with this problem related to the diagnosis accuracy, Computer-Aided Diagnosis systems are often used as a second assistant reader to improve the accuracy of diagnosis made by radiologists during screening practicality.

Computer-aided diagnosis systems are efficient tools that are widely used for Medical Image Analysis to improve diagnosis accuracy [2] [3] [4]. Medical image analysis lies at the basis of these systems. CAD systems, used for medical image analysis, consist of a stepwise process, which is usually designed according to the problem given at hand. Generally, it involves preprocessing, segmentation, detection and classification techniques.

Automated analysis of medical images is an important field in today's world of research. Researchers started working on medical image analysis as soon as they had access to medical image acquisitions on computers. Early in the late 1990s, most of automated analysis systems were based on conventional image processing methods, such as morphological processing [5], edge detection [6], region growing [7] and many more.

The goal of these works was to achieve a rule-based system that solves a particular problem. GOFAI (Good Old Fashioned Artificial Intelligence) or symbolic IA is the name attributed to these systems. The concept of a GOFAI system relies on the idea that cognition can be represented as a sequence of computational terms. So to solve a particular problem related to medical image analysis, all we have to do is to find the right stepwise computational system [8].

Since the late 1990's, Supervised techniques have gained popularity in medical image analysis field [9], and most of today's systems that are built on supervised techniques, particularly those used for commercial purposes, are now very successful. In supervised techniques, such as Active Shape Model (ASM) and Active Appearance Model (AAM) [10], we use data to build the system, and up to now this approach is still widely investigated in actual researches for mainly two reasons: the abundance of public data sets and the availability of computational machines and services with good CPU/GPU performance that is needed to build, train and test the model. Owing to all these improvements, it can be noted that there is a big transition from systems that are based on crafted features to systems that are trained automatically using available datasets.

In the beginning, systems used hand-crafted methods that are designed by humans to extract features from the data to learn. The next level of this approach is to let the system itself extract automatically the features that best represent the data for the problem given at hand. This can be done by transforming the given inputs to labeled outputs while learning increasingly the extraction of high-level features.

Substantial amount of works has been proposed in the literature related to Medical Image Analysis with deep learning approaches [11]. One of the most successful deep learning models that have been widely used in Medical Image Analysis is Convolutional Neural Networks (CNNs) [11]. They saw their first real-world successful application in LeNet [12], which was a model designed by LeCun in 1998 for hand-written digit recognition. CNN became popular in 2012 when a model called AlexNet [13] has been proposed in ImageNet competition. The model won the challenge with a great margin outperforming all competitors. And in the next years that followed, substantial amount of work has been proposed with more enhancement using related architecture.

In medical image analysis, CNNs are one of the best choices made by researchers to design and build efficient CAD systems. Many methods have been used for feature extraction and were very popular before the breakthrough of CNNs in 2012. Examples include Principal Component Analysis (PCA), Sparse Coding approaches, and other techniques that have been well detailed in [14].

With regards to lung nodule classification, CNNs outperform most of classical feature learning methods [15]. The proposed work is inspired by these pivotal developments in Medical image analysis researches that are related to CNNs.

In this paper, we introduce an efficient approach for lung nodule classification based on both CNNs and DCT for representation learning. Only relevant information acquired with Discrete Cosine Transform is fed to our Convolutional Neural Network instead of raw patches that are extracted from CT-images from which features are usually extracted. CNN is then used for feature extraction from the DCT output with Convolution, Max Pooling and Dropout layers as presented on Fig. 2.

The rest of this paper is organized as follows: In material and methods, we give a brief overview of Machine learning concepts, Convolutional Neural Networks, and Discrete Cosine Transform. Then, we describe our contribution related to lung nodule classification combining both Convolutional Neural Network and Discrete Cosine Transform. Finally, we provide all the details of the experimentation performed to evaluate the proposed method, then we discuss obtained results and also open challenges and future works.

II. Material and Methods

A. Machine Learning

Machine learning approaches are divided into two major categories: supervised and unsupervised learning algorithms. In supervised methods, the model is described using a dataset of n entry $x_{i\in\{1,\dots,n\}}$ that is defined as:

$$D = (x_i, y_j) \mid x_i \in I, y_j \in O, i \in \{1,\dots,n\}, j \in \{1,\dots,m\}$$

x_i is the input, y_j is the output or the label of the input associated to y_j and n is the total number of the dataset entries. The output y_j can take several forms according to the problem given at hand. For example in our classification problem, the output can be defined as a scalar of type Boolean: $true$ for "it is" or $false$ for "it is not" a nodule, while in other problems y can be a multi-dimensional vector.

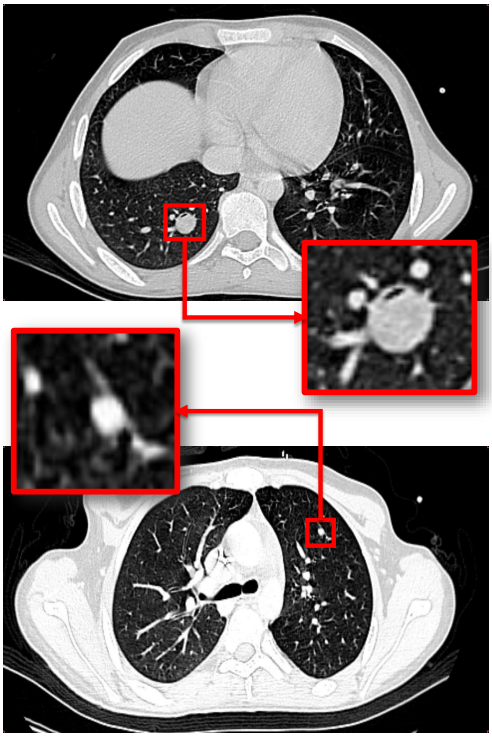


Fig. 1. Example of Big and Small Nodules on Two Different CT Images.

In supervised learning, the model analyzes the data pairs (x, y) that are fed to it, and produces an inferred function $f(x, \theta)$ where x is the input, and θ is the model parameters. This function can be used to map new unseen entries other than the pairs (x, y) used to train the model. The parameters θ are computed based on a loss function $loss(y, y')$ where y' is the label obtained by $f(x, \theta)$.

Differently from supervised learning, an unsupervised model process the input data without any pre-defined labels which help find or discover previously unknown patterns. Examples include Principal Component Analysis and clustering methods. The last one is often used to group the dataset elements into one or multiple groups in such a way that the elements of a given group share similar properties more than other elements in a different group.

Because of these nuances, unsupervised models cannot be applied directly to a classification or regression problem since there is no pre-defined outcome that gives us an idea of what the output should be. Supervised learning approaches are often used in pulmonary nodule classification problem since we want to get a better understanding of the nodule structure so the model can tell if a given patch is a lung nodule or not.

B. Convolutional Neural Networks

Artificial Neural Network (ANN) is an information processing model that lies at the basis of most deep learning methods. A neural network consists of many interconnected neurons just like the biological nervous system but less complicated. A neuron is a node that has many inputs and one output. It

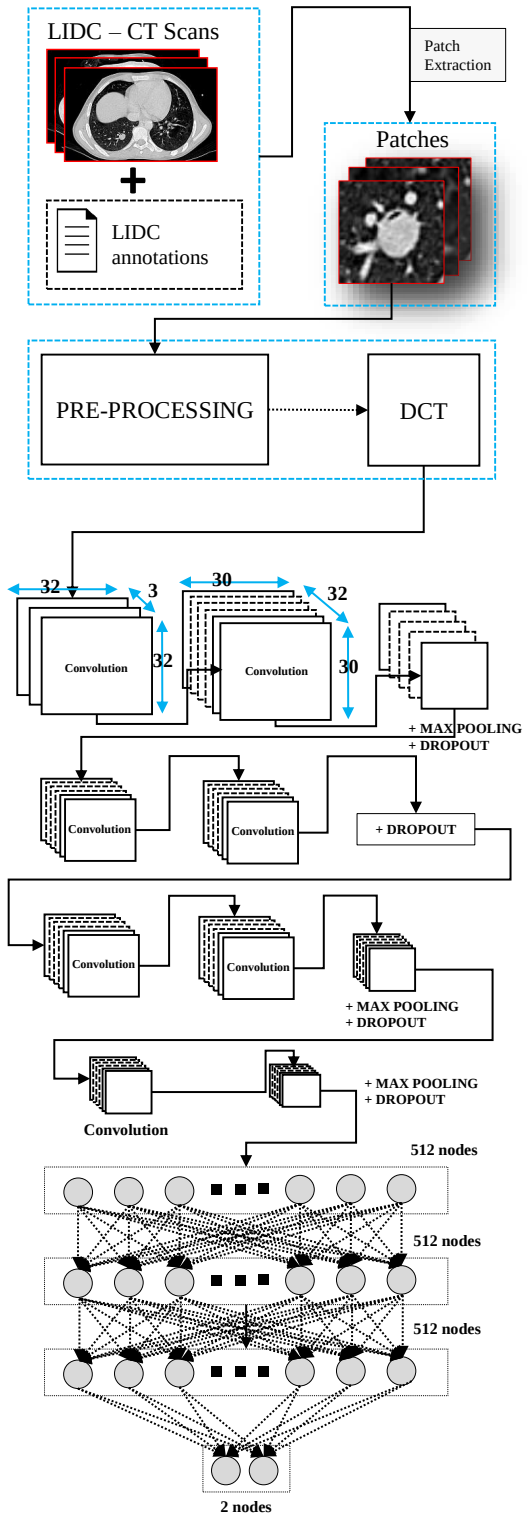


Fig. 2. The Proposed Model Architecture. The Scheme Shows Different Steps from Data Acquisition from LIDC Database to Classification. Also it Shows Different Layers used for Feature Extraction (Convolution, Max Pooling and Dropout) and it Shows also the Classifier Structure.

can be defined as a function that receives an input data x and provides an output y.

The output is the result of an activation function that takes as argument a linear combination of the input x fed to the neuron plus the $bias$ of the neuron. It is defined as:

$$activation = f(bias + \sum_{i=1}^{n} w_i \times x_i)$$

Generally, the activation functions can be divided into two types: linear and nonlinear functions. The last one is often used because it makes it more efficient and easy for the model to adapt to a variety of data. Examples include Sigmoid, Hyperbolic Tangent, Rectified Linear Unit ($ReLU$), and Exponential Linear Unit (ELU).

Convolutional Neural network (CNN) is a conventional Deep Learning model, an improvement in Artificial Neural networks, that is widely used in the field of computer vision. Differently from the MLP (one of the most well-known conventional model of ANN), the CNN involves convolution operation. It contains multiple convolution layers that lie at the basis of its model architecture. With this property at hand, the model performs one training process for the same object occurring at different positions in the different images. In fact, we repeat the following process for each layer: a convolution operation is performed on the input image using a set of kernels and biases W,B, which gives us a feature map X as a result of this process. Next, a nonlinear transformation is performed on the obtained feature. This transformation is defined as follows:

$$X_n = transform(W \times X_{n-1} + bias)$$

Another advantage of CNN is the parameter-dimensionality reduction. Pooling operation is applied to each feature Map to progressively reduce the number of parameters and thus computation complexity of the model. At the end of these two processes, fully connected layers are usually added to the network to complete the model.

C. Discrete Cosine Transform

Discrete Cosine Transform is process that has been introduced by Ahmed et al. [16] in 1974. It is often used in image processing to deal with dimensionality reduction and image compression[17] [18]. It is also used to extract the most relevant information in the image and it is very efficient in a stepwise image processing system, particularly when the DCT coefficients are only used for image representation instead of the whole image.

When the DCT is performed on a raw image, it transforms the image representation from the spatial domain to the frequency domain. Additionally, DCT is data independent due to its fixed basis and it can be used as a simple matrix operation. The DCT formula is defined as:

$$DCT(x,y) = \begin{cases} \frac{2}{M \times N} \sum_{y=0}^{N-1} p(x,y) \times c(x,y) \\ c(x,y) = cos(\frac{(2x+1)*u_n}{2N} \times cos(\frac{(2y+1)*v_n}{2M})) \end{cases}$$

where $DCT(x,y)$ represents the DCT's coefficients and $p(x,y)$ represents the image patch or pixels that will be performed by DCT.

TABLE I. THE PROPOSED MODEL ARCHITECTURE- IT SHOWS ALL THE DIFFERENT TYPES OF LAYERS THAT FORM THE BASIS OF OUR NETWORK. IT SHOWS ALSO THE SHAPE SIZE OF EACH OUTPUT ALONG WITH THE TOTAL NUMBER OF ITS ASSOCIATED PARAMETERS

LAYERS	OUTPUT	TOTAL PR.
2D CONVOLUTION	$32 \times 32 \times 3$	896
2D CONVOLUTION	$30 \times 30 \times 32$	9248
2D MAX POOLING	$15 \times 15 \times 32$	0
DROPOUT	$15 \times 15 \times 32$	0
2D CONVOLUTION	$15 \times 15 \times 32$	18496
2D CONVOLUTION	$13 \times 13 \times 64$	36928
DROPOUT	$6 \times 6 \times 64$	0
2D CONVOLUTION	$6 \times 6 \times 64$	0
2D CONVOLUTION	$6 \times 6 \times 64$	36928
2D MAX POOLING	$4 \times 4 \times 64$	0
DROPOUT	$2 \times 2 \times 64$	0
2D CONVOLUTION	$2 \times 2 \times 64$	36928
2D MAX POOLING	$2 \times 2 \times 64$	0
DROPOUT	$1 \times 1 \times 64$	0
FLATTEN	64	0
DENSE	512	33280
DROPOUT	512	0
DENSE	512	262656
DROPOUT	512	0
DENSE	512	262656
DROPOUT	512	0
DENSE	2	1026

Most of the relevant data that represents the image is concentrated in a few coefficients of the DCT which makes it very efficient in data-dimensionality reduction.

Thus it can used as a first step in the feature extraction process, and instead of using directly the patches extracted from CT-images, we integrate the DCT transformation as a first step to boost the performance of our classification Model.

III. PROPOSED METHOD

In this work, we perform a stepwise classification system for the pulmonary nodule. First, CT-images are transformed from the spatial domain to the frequency domain using DCT. Then, these DCT coefficients (which represent the most relevant information in the images) are fed to the CNN whose architecture is defined as follows:

First of all, we start with the input which has a shape of 32×32. The input is a grayscale $2D$ patch extracted from the full CT-image contained in our dataset. It is extracted based on the pairs (x, y); the nodule location coordinates in 2D CT-slices. All the pairs are provided in one file included in the $LIDC$ database.

Before feeding the input to the CNN, there are two pre-steps: data augmentation and Discrete Cosine Transform. We use data augmentation to improve the diversity of our available dataset.

In this work, the data augmentation technique we are using include translation, rotation and cropping. The output associated with the new input obtained after data augmentation is manually validated by the practitioner. In total, we have 8000 patches that we divided into three subsets: training, testing, and validation. The DCT transform is applied on each patch p_i of the data set $input = DCT(p_i)$ which gives us a new input that will be fed to our network.

As we mentioned before, the DCT transform is used to improve the effectiveness of the classification process by feeding only the most relevant information of the input to the network. The feature extraction comes after the two pre-steps. The model architecture is described in figures on Table 1.

First, a 2D convolution is applied to the input using 32 different filters of size 3×3. The convolution is applied twice: the first one involves padding while the next one doesn't. For each convolution, we use $ReLU$ as an activation function to increase the output non-linearity.

In the next step of the process, we use Max-Pooling after convolution to down-sample the convolution output-representation. In this layer we use a shape of 2×2 which reduces the dimensionality of the convolution output from $30 \times 30 \times 32$ to $15 \times 15 \times 32$.

After the max-pooling comes Dropout. The goal of this layer is to prevent the model from overfitting. It consists of selecting randomly neurons and turns them off during each iteration of the training process. In fact, the dropout layer turns off P of neurons in each iteration, where P is the percentage of neurons to turn off randomly during the training process.

Since convolution layers have few parameters, they require less regularization as a starting point; hence we set the P value to 25% $(P = 25\%)$ for each Dropout layer.

In this work, we perform the process: convolution $\rightarrow$ maxPooling $\rightarrow$ dropout 4 times. We use for the convolution layers different shapes of size: 32×32, 64×64, 64×64 and 64×64 respectively. Also, we use the same activation function $ReLU$ for all convolution layers to improve the non-linearity of their output.

We use Flattening at the end of the convolution process to convert the last output data into a one-dimensional array which will be used as the feature vector. The next part of our model is the Fully Connected Neural Network which consists of 4 Dense layers: the input (512 Nodes), 2 hidden layers (512 Nodes each) and the output (2 Nodes). Again, after each dense layer, we add a dropout layer to prevent the model from overfitting. We use ReLU as the activation function for all layers except the output, for which we use $SoftMax$ as an activation function.

In the next section, we will describe the experimentation we built to evaluate the proposed model. We will describe in detail the database we used and the behavior of our model. Finally, we will report our experimentation results and we give a brief overview of the work perspectives.

IV. EXPERIMENT AND RESULTS

Computer-Aided Diagnosis systems are based generally on the following stepwise processing system: 1) data acquisition, 2) medical image preprocessing, 3) medical image segmentation, 4) detection, and 5) classification or false positive reduction.

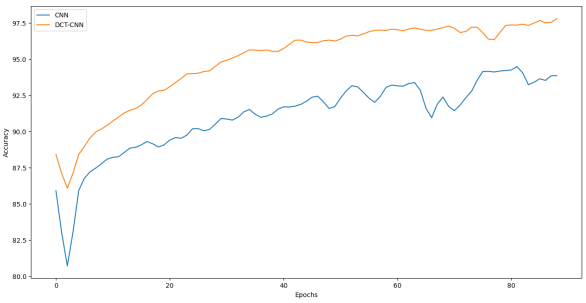


Fig. 3. The Evolution of the Training Accuracy of Both the Proposed Method (DCT + CNN) and the Conventional CNN.

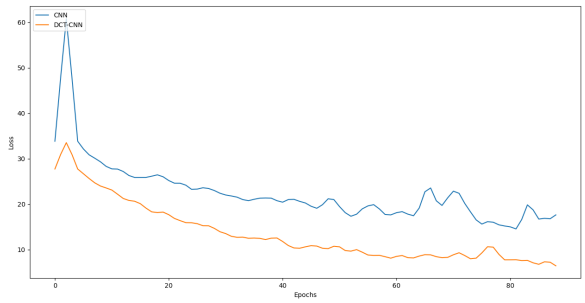


Fig. 4. The Evolution of the Training Loss of Both the Proposed Method (DCT + CNN) and the Conventional CNN.

In this work, we are focusing on the Classification which is the main subject of our research. Other efficient approaches that focus mainly on medical image preprocessing and segmentation are well structured and detailed in these works [19], [20], [21], [22]

The main goal of this work is to evaluate the impact of combining Discrete Cosine Transform and Convolutional Neural Network on the classification accuracy for pulmonary nodules, to determine whether or not the proposed method outperforms the standard CNN classifier. It is also our goal to improve the classification accuracy of the standard CNN model. We do not include in this experiment, comparison between CNN and other Methods since the proposed work aims at improving the Classification accuracy of the standard CNN. Detailed comparison of CNN with the state of the art of Deep learning approaches for medical image analysis are presented in [15].

In this experiment, we use lung CT-images from the well-known LIDC database (Lung Image Database Consortium) [23]. The LIDC is an efficient international web-accessible database that is widely used for development, training, and evaluation of Computer-Assisted Diagnosis systems (CAD) that target lung cancer detection and classification.

Each Lesion is marked-up by multiple experts. The coordinates of the lesion center (x, y) on the CT-image as well as its radius, all are provided on the database to help Medical Image Analysis researchers evaluate easily their built systems.

TABLE II. This Table Shows the Test Accuracy for Both Methods Labeled CNN DCT-CNN(Our Proposed Method). It Shows also Both the Average of Accuracy During all Training Process and the Average of Accuracy after Hitting the Max Accuracy Until the End of the Training Process (EOT). The Average Loss is also Depicted along with the Average Loss after Hitting the Min Loss value Until the End of the Training Process.

Methods	**CNN**	**DCT-CNN**
Test Accuracy	*91,78%*	*96,51%*
Accuracy (average)	91.19%	95.10%
Accuracy [max $\rightarrow$ EOT]	93.38%	97.80%
Loss (average)	22.37%	13.02%
Loss [min $\rightarrow$ EOT]	18.51%	06.14%

In this work, we use the center coordinates to extract patches from CT images. The figure 5 shows an example of different patches used to train and test the model.

In total, we have 8000 patches with a shape size of 32×32. We divide the obtained patches into 3 subsets: training, testing, and validation. The first subset entries are used by our model as labeled examples to learn from. The second subset is used to check the model performance while tuning its hyper-parameters during the training process. Finally, the third subset is used to evaluate the final model fit.

In machine learning, an epoch is a measure that represents the number of times all the training vectors are used once to tune the model hyper-parameters.

In this experiment we are setting its value to $epoch = 15$. The batch is the number of samples passed simultaneously during the training process before the weights getting updated, and this per one epoch. In this experiment, we set the value of $batch = 32$.

For all the layers, we use $ReLU$ as activation function except the output where we use $SoftMax$ as activation function. The number of filters per each convolution layer is 32, 64, 64 respectively, of the same size: 3×3.

In Max-Pooling, we use a 2×2 box and during all Dropout operations we turn off 25% of the neurons which all are chosen at random.

Fig. 3 and 4, we show a graph that consists of two different curves: the blue one which represents the evolution of the classification accuracy/loss after each epoch of our proposed model while the orange curve represents the evolution of the classification accuracy/loss of the conventional CNN.

Fig. 3 represents the evolution of the classification accuracy of the two models: the conventional CNN and our proposed Model. After each epoch we evaluate the classification accuracy of both models using entries from the third subset of LIDC database - Entries that we use only for testing and which we don't use in the training process, to ensure the effectiveness of the testing process.

From Fig. 3 and 4 we can see that the proposed method outperforms the conventional CNN in terms of Accuracy with over 4.73%. In Table II we provided more details about the

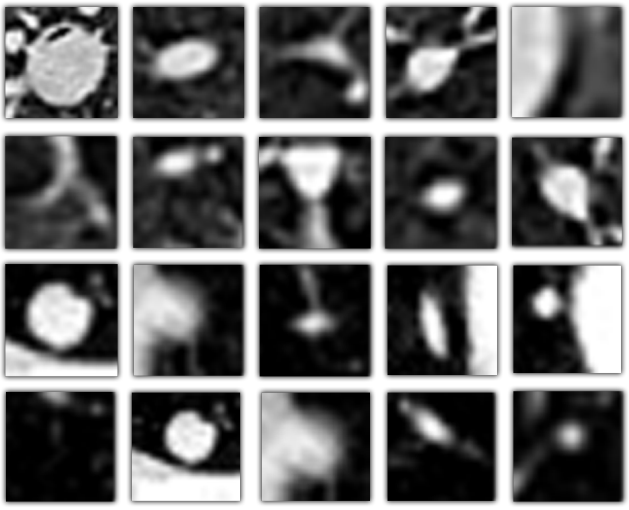


Fig. 5. Example of Big and Small Nodules on Two Different CT Images.

experimental results. It shows the Test Accuracy of both CNN and the proposed Method. It also shows the average accuracy starting from the moment the model reached its maximum accuracy until the end of Training.

We included also values of the classification accuracy from the 61st epoch to the 85th epoch of the experimentation. From both the graphs and the results table we can see good improvement in terms of the classification accuracy when using Discrete Cosine Transform along with Convolutional Neural Network as it refines the information of entries used for training to improve the model accuracy. The final result show that the proposed method outperforms the conventional CNN with a good margin.

V. Conclusion

The main goal of this work was to evaluate the impact of Discrete Cosine Transform (DCT) on the classification accuracy when it's applied along with Convolutional Neural Network (CNN) for Lung Nodules classification.

The proposed method aims at using DCT to extract only most relevant information in the patches before feeding them to the model as a training data. The model architecture, which is also considered a keystone of the model accuracy, is also described in details along with all its parameters.

The proposed Model is tested on LIDC database which is one of the most efficient datasets used for lung nodules classification and detection. The proposed Method outperform the standard CNN in terms of accuracy with a good margin.

In this work, we demonstrated that Discrete Cosine Transform can improve the accuracy of the conventional CNN with a good margin (in our experiment: between 4.73%), when it is applied for Lung nodules classification in CT-images. In future works, this proposed method can be used as the last step that completes a CAD system; a Real-World Application that aims at analyzing each lesion in an input CT-image and could tell if it is a lung nodule or not.

References

[1] K. D. Miller, L. Nogueira, A. B. Mariotto, J. H. Rowland, K. R. Yabroff, C. M. Alfano, A. Jemal, J. L. Kramer, and R. L. Siegel, "Cancer treatment and survivorship statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 5, pp. 363–385, 2019.

[2] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (hpbcr) using optimized ensemble learning," *Computational and structural biotechnology journal*, vol. 15, pp. 75–85, 2017.

[3] M. A. Al-Antari, M. A. Al-Masni, M.-T. Choi, S.-M. Han, and T.-S. Kim, "A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification," *International journal of medical informatics*, vol. 117, pp. 44–54, 2018.

[4] E. Y. Jeong, H. L. Kim, E. J. Ha, S. Y. Park, Y. J. Cho, and M. Han, "Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators," *European radiology*, vol. 29, no. 4, pp. 1978–1985, 2019.

[5] J. Serra and P. Soille, *Mathematical morphology and its applications to image processing*, vol. 2. Springer Science & Business Media, 2012.

[6] S. Chakraborty, M. Roy, and S. Hore, "A study on different edge detection techniques in digital image processing," in *Feature Detectors and Motion Detection in Video Processing*, pp. 100–122, IGI Global, 2017.

[7] M. Dabass, S. Vashisth, and R. Vig, "Effectiveness of region growing based segmentation technique for various medical images-a study," in *International Conference on Recent Developments in Science, Engineering and Technology*, pp. 234–259, Springer, 2017.

[8] B. C. Smith, *The promise of artificial intelligence: reckoning and judgment*. Mit Press, 2019.

[9] M. W. Berry, A. Mohamed, and B. W. Yap, *Supervised and Unsupervised Learning for Data Science*. Springer, 2019.

[10] M. Iqtait, F. Mohamad, and M. Mamat, "Feature extraction for face recognition via active shape model (asm) and active appearance model (aam)," in *IOP Conference Series: Materials Science and Engineering*, vol. 332, p. 012032, IOP Publishing, 2018.

[11] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[14] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[16] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[17] N. Ponomarenko, V. Lukin, K. Egiazarian, and J. Astola, "Dct based high quality image compression," in *Scandinavian Conference on Image Analysis*, pp. 1177–1185, Springer, 2005.

[18] M. Sun, X. He, S. Xiong, C. Ren, and X. Li, "Reduction of jpeg compression artifacts based on dct coefficients prediction," *Neurocomputing*, vol. 384, pp. 335–345, 2020.

[19] A. El Hassani and A. Majda, "Efficient image denoising method based on mathematical morphology reconstruction and the non-local means filter for the mri of the head," in *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pp. 422–427, IEEE, 2016.

[20] A. El Hassani, B. A. Skourt, and A. Majda, "Efficient lung ct image segmentation using mathematical morphology and the region growing algorithm," in *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, pp. 1–6, IEEE, 2019.

[21] B. A. Skourt, A. El Hassani, and A. Majda, "Lung ct image segmentation using deep neural networks," *Procedia Computer Science*, vol. 127, pp. 109–113, 2018.

[22] A. Majda and A. El Hassani, "Graph cuts segmentation approach using a patch-based similarity measure applied for interactive ct lung image segmentation," *International Journal of Computer and Information Engineering*, vol. 12, no. 7, pp. 520–524, 2018.

[23] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

Streaming of Global Navigation Satellite System Data from the Global System of Navigation

Liliana Ibeth Barbosa-Santillán[1]
Computer Science Department
University of Guadalajara
Guadalajara, México

Juan Jaime Sánchez-Escobar[2]
Research department
Technical and Industrial Teaching Center
Guadalajara, México

Luis Francisco Barbosa-Santillán[3]
Mechatronics Department
University Technological of Puebla
Puebla, México

Amilcar Meneses-Viveros[4]
Computer Science Department
CINVESTAV,México City

Zhan Gao[5]
School of Computer Science and Technology
Nantong University, Nantong, China

Julio César Roa-Gil[6]
Computer Science
ITESM, Guadalajara, México

Gabriel A. León-Paredes[7]
GIHP4C Research Group
Universidad Politécnica Salesiana
Cuenca, Ecuador

***Abstract*—The Big Data phenomenon has driven a revolution in data and has provided competitive advantages in business and science domains through data analysis. By Big Data, we mean the large volumes of information generated at high speeds from various information sources, including social networks, sensors for multiple devices, and satellites. One of the main problems in real applications is the extraction of accurate information from large volumes of unstructured data in the streaming process. Here, we extract information from data obtained from the GLONASS satellite navigation system. The knowledge acquired in the discovery of geolocation of an object has been essential to the satellite systems. However, many of these findings have suffered changes as error vocalizations and many data. The Global Navigation Satellite System (GNSS) combines several existing navigation and geospatial positioning systems, including the Global Positioning System, GLONASS, and Galileo. We focus on GLONASS because it has a constellation with 31 satellites. Our research's difficulties are: (a) to handle the amount of data that GLONASS produces efficiently and (b) to accelerate data pipeline with parallelization and dynamic access to data because these have only structured one part. This work's main contribution is the Streaming of GNSS Data from the GLONASS Satellite Navigation System for GNSS data processing and dynamic management of meta-data. We achieve a three-fold improvement in performance when the program is running with 8 and 10 threads.**

Keywords*—GLONASS; streaming; extraction; satellites data; observation files; metadata

I. Introduction

The data collection does not present a problem. However, handling these volumes of information poses a challenge to the industry. The fundamental challenge regarding large volumes of data from different sources is identifying new uses that have not been found. Companies' challenge is to develop methods of realizing the real value of this mine of terabytes of data. Big Data is the medium through which these large volumes of information acquire significant value [1] [2].

Many types of problems can occur during the different stages or processes involving Big Data. For example, in the transformation step, the storage process and the extraction process in streaming from other sources, there are significant challenges related to linking variables such as the speed, volume, and variety of data extracted and processed. Similar effects arise in the preprocessing of the data. The hardware capability plays a vital role in a system's ability to clean data in the shortest possible time. When analyzing and visualizing information, it should be presented in the easiest and simplest possible way so that anyone can understand it. One challenge in this vein is the use of techniques and methodologies that summarize and display information clearly and accurately [2] [3] [4].

The Global Navigation Satellite System (GNSS) combines several existing systems for navigation and geospatial positioning, including the Global Positioning System (GPS), GLONASS (Global System of Navigation), and Galileo (a European radio-navigation program) [5] [6].

GPS was the first system GNSS. It was released at the end of 1970 by the Department of Defense of the United States; it uses a constellation between 24 and 32 satellites and provides global coverage. The Ministry of Defense of the Russian Federation operates GLONASS; this consists of 31 satellites. Twenty-four are active; three are in backup, two in maintenance, and two more in testing. GALILEO is the European radio navigation and satellite-positioning program developed by the European Union in conjunction with the European Space Agency and expected to be officially available for civil use by 2020. In November 2016, four new satellites launched, giving 18 satellites already in orbit. These systems are composed of Space-Based Augmentation Systems (SBAS) or Ground-Based Augmentation Systems (GBAS). Examples of SBASs are the US-based Wide-Area Augmentation System (WAAS), the European Geostationary Navigation Overlay Service (EGNOS), and the Japanese Multi-functional Transport Satellite (MTSAT) based on SBAS. The GNSS signal radio is involved with frequencies close to 1.5 GHz (1.5 billion

cycles per second). GNSS signals operate at higher frequencies than FM radio signals but lower than those of a microwave oven; when GNSS signals reach land, they are fragile. An electromagnetic wave arriving from space must traverse three distinct zones before going a receiver on the Earth's surface: the vacuum, the ionosphere, and the troposphere. The signal delay increases with the propagation of time.

It arises from two factors: the propagation speed and the increase in the trajectory's length due to bending by refraction. In the vacuum, this delay is negligible, and is the programming time proportional to the distance depending on the light, whatever the frequency of the wave is. In the ionosphere (at altitudes of 100 to 1000 Km), ultraviolet, solar, and other radiation types, ionize gaseous molecules and release electrons. The number of free electrons per cubic meter varies between 10-16 and 10-19. The delay is proportional to the number of free electrons encountered by the signal along its path and is dependent on the inverse of the square of wave frequency. It varies for each particular point, according to its latitude, direction, and observation moment. The delay may change in the zenith by between 2 ns. and 50 ns. For frequencies in the L-band, the delay can reach up to 2.5 the factor due to the trajectory's inclination. Its effect at midday is up to five times between midnight and dawn. The last area that the wave traverses is the troposphere and the other regions of the upper atmosphere. Although this area extends to heights of up to 80 km, significant delays are incurred only in the lower 40 km. This delay corresponds to increments in the distance of the order between 1 m at the zenith and up to 30 m and five elevation grades (advanced GPS). GNSS systems continuously transmit signals at two or more frequencies within the L band. These signals contain range codes and navigation data. The main components of the signal are:

- A carrier, a sinusoidal radio signal, is a specific frequency.
- A ranging code consists of sequences of zeros and ones that allow the receiver to determine the Satellite radio signal's travel time to the receiver. These are called PRN sequences or PRN codes.
- Navigation data consists of a message that provides information on the satellite's ephemeris (pseudo- Keplerian Elements or the satellite's position and velocity), clock parameters, and error margins (a set of low-precision ephemeris data), the satellite status, and additional information.

Frequency-band mapping is a complex process since multiple users and services can access the same range. In other words, the same frequencies are for different purposes in different countries. The International Telecommunication Union (ITU) is a United Nations agency that coordinates the radio spectrum's shared global use. ITU divides the electromagnetic spectrum into frequency bands, with different radio services assigned to particular bands. Two band segments are given to the Aeronautical Radio Navigation Service (ARNS) at the primary level world- wide. These bands are for the safety of life (SoL) applications, and no other use of these bands can interfere with GNSS signals. These segments are the upper L band (containing the GPS bands $L1$, Galileo $E1$, GLONASS $B1$, and Beidou L), and the lower L band (including the $L5$ band of GPs, $G3$ of GLONASS, $E5$ of Galileo, and B2 of Beidou). Receiver Independent Exchange Format (RINEX) was developed by the Institute of Astronomy at the University of Bern to enable the exchange of the GPS data collected during the Europe Reference Frame (EUREF), which included more than 60 GPS receivers from four different manufacturers. In the development of this format, it was taken into account that most software for geodesic processes for GPS data use a well-defined set of observables, including:

- Measurement of the carrier signal phase is a measure of the receiver's satellite carrier signal frequency, as shown in Equation 1.

$$\begin{aligned} Phase(tight) = Phase(r) \\ -RealTime(r) \times frequency \end{aligned} \tag{1}$$

  where: r = clock

- Measurement of the pseudo-range is the difference in the reception time (expressed in the receiver's time frame) and the transmission time (described in the satellite's temporary framework) of a different satellite signal.

$$\begin{aligned} PR = distance + c \times Shiftingreceiverclock \\ +Satelliteclockshifts + Otherbiases \end{aligned} \tag{2}$$

  where: PR = PseudoRange and c = cycles

- The observation time is reading the receiver's clock at the moment of the phase carrier's validity and code measurements.

Version 3 of the RINEX format consists of three types of ASCII files: an observation data file, a navigation message file, and a meteorological data file [7].

Each file type contains a header section and a data section. The header section contains global information for the file at the beginning of it. This header section contains labels in columns 61-80 for each line in the area; these tags are mandatory and must appear as required by the format. RINEX requires a minimum amount of space, regardless of the number of different observables, the specific receiver used, or the satellite system. It indicates in the heading the types of observations recorded by each receiver and satellite system observed. There is not a maximum length for each record to limit these observations. Each meteorological data and observation file contains data from a site and a session. In RINEX version 3, navigation message files can contain messages from more than one satellite system (e.g., GPS, GLONASS, Galileo, or SBAS). GNSS observables require two fundamental quantities to be defined: The time and phase.

The timeof measurement is the time recorded by the receiver of the signals. It is similar for phase and range measurements and is identical for all satellites observed. For single system data files, expressed by default in the respective satellite's time system: otherwise, the actual time (for mixed files) is in the start time header log.

Phase involves the carrier wave and its complete cycle measures. The semi-cycles measured by quadrant-type receivers must be converted into full cycles and marked with the respective observation code—the phase changes in the same direction as

the range (a negative Doppler effect). Observable ones are incorrect for external influences such as atmospheric refraction and satellite offsets. Phase changes between phases of the same frequency but tracked in a different carrier channel are not corrected.

The knowledge acquired in the discovery of geolocation of an object has been essential to the satellite systems. However, many of these findings have suffered changes in error localization and many data. The Global Navigation Satellite System (GNSS) combines several existing navigation and geospatial positioning systems, including the Global Positioning System, GLONASS, and Galileo [8]. We focus on GLONASS because it has a constellation with 31 satellites.

The motivation of the proposed work is to extract information in real-time based on the Glonass positioning system. Research gaps consist of the GLONASS navigation file defines the orbits of the satellites by their coordinates inserted from the central bases at certain times and indicating the age of said information.

The definition of GLONASS time has also given its problems, being necessary to indicate the origin of the observations' reference time. On average, the navigation files consist of 150 lines and the observation files of 33,500 lines. The RINEX observation files could contain a receiver-derived clock offset.

The data (epoch, pseudo interval, phase) have been previously corrected or not for the reported clock shift. RINEX Versions 2.10 onwards requests a clarifying header record: RCV CLOCK OFFS APPL. Then it would be possible to reconstruct the original observations, if necessary.

Our research's difficulties are (a) To handle the amount of data that GLONASS produces efficiently and (b) to accelerate data pipeline with parallelization and dynamic access to data because these have only structured one part. This work's main contribution is the Streaming of GNSS Data from the GLONASS Satellite Navigation System for GNSS data streaming processing and dynamic management of metadata implemented within the database. We achieve a three-fold improvement in performance when running the program with 8 and 10 threads. Our research questions are as follows:

P1 Is it possible to automatically identify and download RINEX data sources from GLONASS?

P2 How can RINEX files be identified on semantics?

Our research hypothesis is as follows:

H1 It is possible to discover RINEX files in GLONASS based on semantics.

The paper is structured as follows: Section 1 a brief theoretical framework. Section 2 describes related work. Section 3 discusses stream extraction of GNSS data based on the GLONASS satellite navigation system. Sections 4 and 5 present the details of our data sets, evaluation metrics, and our results. Finally, Section 6 presents the conclusions.

II. Related Work

Several works related to this research are:

- GLONASS data stream processing,
- a parallelization mechanism for the ETL module, and
- Managing dynamic structures in a database for Big Data tasks in GLONASS for data mining and satellite data processing.

From the perspective of stream processing in GLONASS, prior works have focused on positioning and kinematic processing through GPS. Some examples of these are studied by Li et al. [9], Wang et al. [10], and Rieke Matthes et al. [11]. Some of these works have been applied to atmospheric measurements [12], [13], in which the main idea is to provide accurate positioning in real-time with minimum error. These systems have evolved to establish services at the cloud computing level, as reported by Karimi et al. [14] and Liu et al. [15]. Several approaches are to optimize the ETL module. These works include an optimization involving the environment (a distributed system), a dynamic design, and the ETL module's parallelization. The ETL module's optimization process allows processing times and efficiently designing the data warehouse structure. In [16] the authors combine GPS, GLONASS and Galileo in order to obtain precise points positioning in real-time. In [17], Koyptov et al. present a system based on data stream collection and processing to determine the geographic coordinates of Earth's ionospheric regions. Kakooei and Tabatabaei [18] develop a hybrid-heterogeneous parallel GPS acquisition algorithm working with a GPU and a multicore CPU.

Distributed parallel architectures proposed in which the ETL module works in the stream with large data volumes. An example of this is the works of Agrawal et al. [19], Boja et al. [20], Ding et al. [21], who focus on the distributed file systems and the ETL module operating in the distributed environment. Bala et al. [22] present a distribution model in order to get fine-grained data for the ETL process.

There are several works on the parallelization of the ETL module. Xiufeng et al. [23], and Radonić et al. [24] present a programming framework that uses Map-Reduce to achieve scalability. This framework is called ETLMR, and it is on a data warehouse; it constructs star schemes and snowflakes and works in dimensions that change over time. This work is evolving to include cloud computing support through the CloudETL framework [25]. In Bala et al. In [26], develop P-ETL, an ETL module that operates on a data warehouse. It runs in parallel in a cluster under the MapReduce paradigm. Masouleh et al. [27] develop an optimization for the execution time of the ETL module using parallelization methods in the shared memory cache of the distributed system. Thomsen et al. [28] propose a framework that allows the parallelization of the ETL module in terms of the three phases that it involves (extraction, transformation, and load). The framework parallelize s the task level as the data level, depending on the stage worked on. This framework works on a node with a multicore processor. Diouf et al. [29] give a review of several speedup ETL process methods.

The use of dynamic structures in the database is implemented in the last few years; the main idea is to handle various types of data transparently. In Big Data, this can help by adapting the stream's structure to incorporate the data sent

for analysis. Han et al. [30] survey the different database technologies for handling large volumes of data and high-performance techniques for cloud computing tasks in real-time. Wu et al. [31] examine the fundamentals of the dynamic characteristics required of a data model for Big-Data tasks. Ji et al., [32] and Fan et al. [33] present a review of the Big Data schemes of the different phases, emphasizing the use of dynamic schema in the database.

Currently, the vast amount of information stored in organizations in all market sectors represents a potential source of knowledge that can be explored and extracted. Positioning satellites is a significant source of information for various companies, especially those working on research geospatial data [34].

Satellites accumulate records in the form of telemetry data points. The telemetry frames format records several hundreds of thousands of data points at each time step, stored in knowledge bases for analysis. These points formed the basis of accumulated historical data and were extracted to give relevant information. Evidence of this includes the study of telemetry satellite data [35] using data mining processes, identifying and categorizing the parameters carried out within the data warehouse, where the data are prepared (standardized and reformatted). After processing, these data are metadata tags, through which experts and users can find categorized information of interest.

In satellite data applications, we know that telemetry data are the only source from which to identify and predict anomalies in artificial satellites. Although there are people who specialize in analyzing these data in real-time, these datasets' large size makes this analysis extremely difficult. Therefore, clustering algorithms are applied to help traders and analysts perform the task of analyzing the telemetry data [36].

Two real cases of anomalies in satellites on space missions are in Brazil. It was possible to evaluate and compare the effectiveness of the two clustering algorithms of K-means and Expectation-Maximization (EM). Their effectiveness was in several telemetry channels, which tended to include outliers; in these cases, they could support satellite operators, allowing for the anticipation of anomalies. However, for silent problems, in which there was only a small variation in a single channel, the algorithms were less efficient. Current cyclone detection techniques and monitoring through models and field measurements do not provide truly global coverage, unlike remote satellite observations. However, it is impractical to use a single satellite orbit to continuously detect and monitor these events due to the limited spatial and temporal coverage.

One solution to alleviate this problem is to use data sensors on multiple orbiting satellites. This approach addresses the unique challenges associated with knowledge discovery and mining of heterogeneous data stream satellites. It consists of two main components [37]: feature extraction from each sensor measurement to discover a set of cyclones, and knowledge sharing between the different remote sensor measurements, based on a linear Kalman filter, to track the predicted storms. Experimental results using historical hurricane data have demonstrated this approach's superior performance compared to other works. Other satellite television broadcasting applications have also been shown [38] (TV broadcasting). An improved algorithm has to identify the most frequent episodes over the broadcasting-satellite service. Frequent episodes at a specific scale of the alarm data extract to summarize the models obtained. Spatial data mining involves extracting implicit knowledge, spatial relations, or other patterns not stored in explicit form in the spatial databases. Based on this approach, the focus of spatial data mining is on deriving information from spatial datasets.

The geographic coordinates of the "hot spots" in the forest fire regions, extracted from satellite images, are studied and used to detect possible points or locations of fires [39]. It found that these applications may give false alarms. Thus, by comparing the brightness detected in several bands, this false information can be identified, and clustering and Hough transform are used to identify regular patterns in access points and applications classified as false alarms. This implementation demonstrates a spatial data mining application to reduce false alarms based on the set of points obtained from the images. Finally, it considers a data analysis based on data from positioning satellites. This work [40] develops an analytical real-time distributed environment in which analysis and simulation are closely coupled, integrating high-performance implementations of image mining run on dedicated servers. It was possible to simulate earthquakes at both the micro and macro levels based on images (Imageodesy) and historical data.

The header registers report the orientation of the antenna's zero direction and the direction of its vertical axis (hole view) if it is mounted and tilted at a fixed station. Header records can also be used for vehicle antennas. The comparison with other systems is closed since the manufacturers have commercial interests. However, the RINEX file is proven to arrive complete with INEX Viewer and RTKLIB and selected these tools because they are freely accessible.

III. Stream Extraction of GNSS Data based on the GLONASS Satellite Navigation System

In this section, the proposed framework for carrying out the stream analysis of data and its respective architecture are detailed.

The system's overall architecture for transfer and extraction of GNSS knowledge is into four main layers: external components, communication, software, and storage, as shown in Fig. 1.

- **External components**: These are all the elements of the physical system, such as GLONASS satellites, receiving antennas, control stations, and data broadcasters.
- **Communication**: This layer allows the transmission of data streams through the network.
- **Software**: consists of all software elements used and developed to extract, process, and store data.
- **Storage**: This layer consists of a logical meta-model and database, in which relevant information saves from downloaded data.

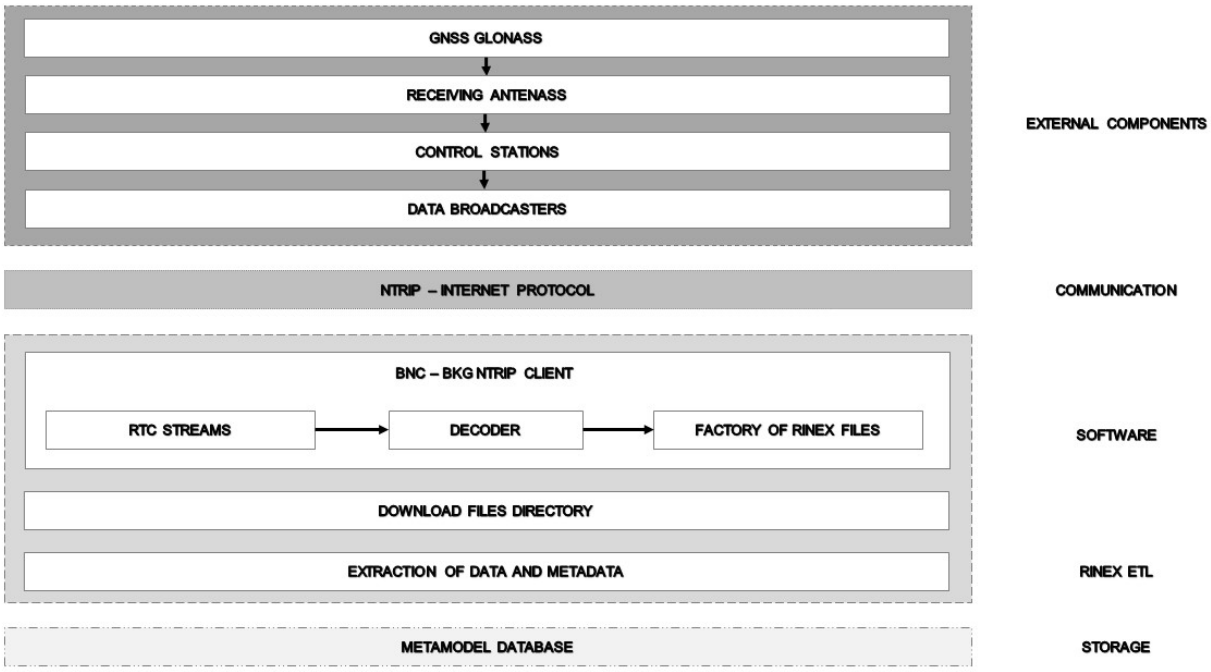


Fig. 1. Stream Extraction of GNSS Data based on the GLONASS Satellite Navigation Architecture.

1) External components: The elements of this layer of the architecture are below.

- **GLONASS:** This is a GNSS constellation with 31 satellites (24 active, three spares, two in maintenance, one in service, and one is undergoing testing) located in three orbital planes with eight satellites each. They produced the primary source from which the data in the RINEX files.
- **Receiving antennas:** These are antennas that receive signals from GLONASS data. They act as an intermediary between GLONASS communication and control stations.
- **Data control stations:** These are the stations in charge of the operation, control, and monitoring of GNSS, and data transmitted by GLONASS are stored here. Information is exchanged with GNSS if a synchronization or reconfiguration event then it is performed for any satellite.
- **Agencies/data broadcasters:** These are agencies, organizations, or institutions that collect, store, process, and investigate the data sent by any GNSS, and in turn relay this data over the Internet to anyone interested in the scientific investigation and processing of this information. The transmission of data takes place over civil frequencies that are open to the public.

2) Communication: This component involves the NTRIP protocol, which allows the transmission of the data streams generated by GNSS over the Internet to client software that receives the information.

3) Software components:

- **BKG Ntrip Client (BNC):** is one of the essential elements, and is a program that simultaneously retrieves, decodes, converts, and processes the data stream from any GNSS system in real-time. It also has some post-processing functions for the RINEX or SP3 files generated by the application. This client software is composed of three main elements, which are:
  - ***RTCM Streams*:** These are the main inputs and consist of data streams downloaded from the agencies or organizations that belong to different networks such as the International GNSS Service (IGS). These data streams arrive in RTCM format.
  - **Decoder:** This element's function is to decode the data streams arriving in the RTCM format and transform it to RINEX version 2.11.
  - **RINEX File Generator:** Once the decoder has completed the transformation, it is responsible for storing the RINEX files in the specified directory.
- **File Directory:** This is the backbone of the storage process for the client software. A well-defined structure is necessary to distinguish between the different types of files generated.
- **Extraction, Transformation, Load (ETL) Tool:**

TABLE I. GENERAL SPECIFICATIONS FOR TESTING

Item	Specifications
Number of files to process	40775
Size of data	80.9 Gigabytes
Maximum error percentage	2
Number of executions	3

TABLE II. GENERAL METRICS FOR TESTING

Metric	Specification
Number of files processed successfully	natural number
Number of processed files flawed	natural number
Execution time	seconds, minutes and hours

A developed specific software application to fulfill the purpose of extracting and transforming data. The data loaded into the database. This application is called RINEX ETL. This application's primary goal is to read the Observation Files and carry out the database's objective data's removal and insertion. The application can run in either serial or parallel mode using the processors in each core of the multiprocessor.

- **The Parser:** is the primary layer of the application and is responsible for reading, extracting, and transforming data from the Observation files. This layer implements the *Runnable* Containers interface, which enables parallel processing through the use of *threads*.
- ***Data Access Object (DAO)*:** This layer provides the standard interface between the application and the database (storage component), allowing communication between the storage and application components.

4) Storage: This section describes the database with a meta-model that is defined to store and retrieve relevant information.

IV. EXPERIMENTS

The objective of the experiments is to analyze and describe the meaning of RINEX files obtained from GLONASS. First, the specifications that apply to all the tests performed on the downloaded files' data are conducted. These specifications are listed in Table I.

The metrics used in the development of the experiments are in Table II.

The tests performed on the data were in four stages: sequential, parallel with multiprocessors, clustering, and queries.

1) **Sequential:** The RINEX ETL application was run in serial mode or with a single thread of execution.
2) **Parallel with multiprocessors:** The RINEX ETL application runs in parallel mode by using threads, making use of the multiprocessor cores. This test was performed with: 8 *Threads*, 10 *Threads*, 16 *Threads*, 32 *threads*, and 50 *Threads*.
3) ***Clustering*:** The application is linked to the database defined with the meta-model. After the database query process, the data loaded into the data mining tool. The clustering process to identify abnormalities in the LLI or find any patterns in the downloaded data from which we could infer and interpret possible improvements. The K-Means algorithm was applied because it is highly parallelizable. K-means was with four clusters.
4) **Queries:** Through the semantics, the following aims will be achieved:
   - Identification of the number of failures in the *Epoch Dates*.
   - Identification of the number of possible cycles slips for observation type $L1$.
   - Identification of the number of possible cycles slips for observation type $L2$.
   - Determination of the frequency of possible cycle slips for observation type $L1$.
   - Determination of the frequency of possible cycle slips for observation type $L2$.

The Rinex file observations are displayed, and it is possible to select one by one the observed satellite constellation in the measurements view, as shown in Fig. 2. If at any point an N appears, it means that the satellite is not observed.

V. RESULTS

This section presents the above experiments' results; we first describe the downloaded files in our sample and then report on the experimental results.
The distribution of the different files downloaded through the BNC-NTRIP client software is in Table III. The total size of the downloaded files was 95.5 GB.

Table IV shows the distribution of Observation Files between correct or readable files and corrupt or unreadable ones due to network or communication problems between the client and the broadcasters or electrical failures on the client-side.

Less than 1% of the files obtained were unreadable and excluded from the experiments. It was notable that at the end of the download time for the data, a longer run time was when implementing the application to process more data to reach a sizeable sample of data in less time.

TABLE III. DISTRIBUTION OF DOWNLOADED FILES

File types	Size (MB)
Observation Files	80,900
Ephemerids Files	105
Raw Data	14,500
Log Files	4,25

TABLE IV. DISTRIBUTION OF THE *Observation Files*

Item	Quantity	Percentage
Total files	40,792	100
Total correct files	40,775	99.96
Total corrupted files	17	0.04

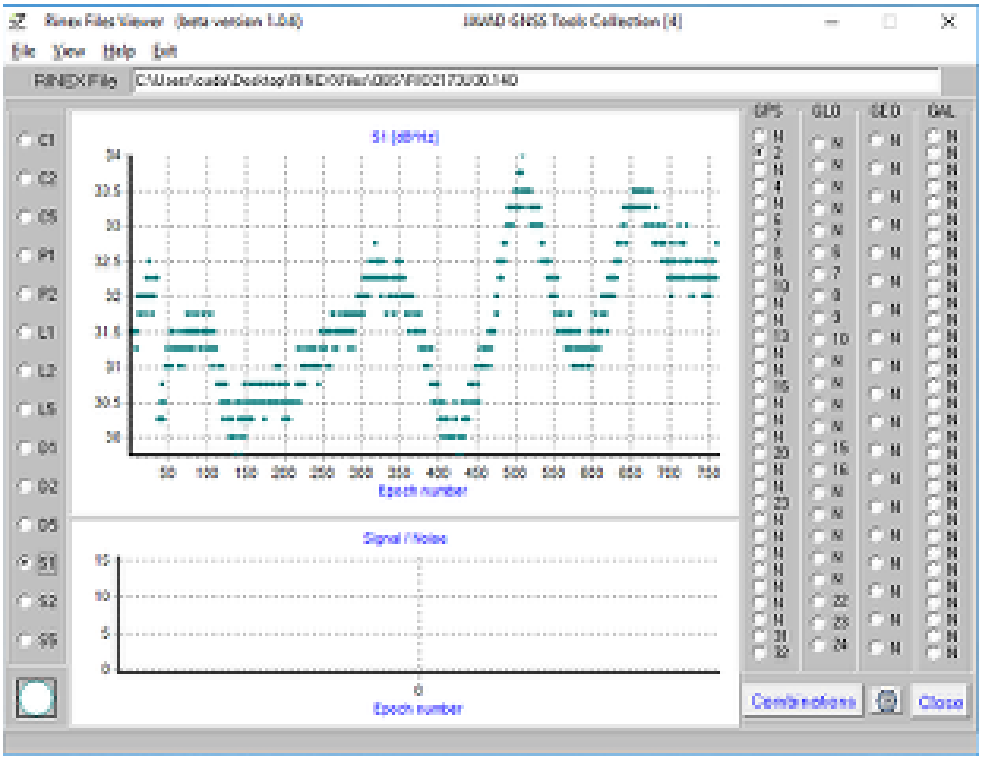


Fig. 2. Measurements view the Observations of the Rinex File.

The maximum number of observations with possible cycle slippage was 5,107,444 for $L1$ and 1,105,203 for $L2$, while the minimum values were 199,511 and 46,604 for $L1$ and $L2$, respectively.

Based on this information, we can infer that one of the more common abnormalities in the $L1$ and $L2$ types of observations is cycle sliding. It is not a failure of the observed value and merely tells us that we must perform a correction process to find the correct value observed. These corrections made using various methods that do not form part of this research.

The results are obtained in the various tests carried out here. The execution time results are described and graphically illustrated for different performance metrics such as CPU usage and memory. The container application, a tool that provides detailed information on containers running on the Virtual Machine (VM), was used to determine and monitor the different metrics.

Table V shows the number of rows saved into the METADATA database, OBSERVATION DATA, and LLI tables of the meta-model.

TABLE V. DISTRIBUTION OF RECORDS IN THE TABLES OF THE META-MODEL

Table	Quantity of records
METADATA	40,775
OBSERVATION_DATA	34,200,319
LLI	12,089,001

TABLE VI. POSSIBLE NUMBERS OF RECORDS WITH CYCLE SLIP, L1 AND L2

Observation type	Quantity of records
L1	2,517,286
L2	1,431,922

Table VI shows the number of records relating to LLI with a possible cycle slip in the $L1$ and $L2$ observation types stored in the LLI table. Within these records, no faults occurred between the epoch dates. It means that no event would alter the observed value when taking the time stamp from GNSS.

The test results for the clustering of values in the observation files for days in which a cycle slip occurred in the $L1$ and $L2$ observations types were as follows: 3,874,181 instances; 28 iterations; and 227.85 seconds to build the model.

Table VII details the epoch date at each of the available satellites according to the cycle's number on $L1$ and $L2$, respectively.

The values reported here correspond to the average of the observed values, as the percentage of instances. For example, there are 36 clustered instances in Cluster 2, as shown in Table VIII.

Table IX shows the clustering test results, in which the data grouped by days. The results were as follows: number of instances: 134, number of iterations: 5, and time is taken to build the model: 0.02 seconds.

Table X shows the percentage of clustered instances. For

TABLE VII. RESULTS, CLUSTER CENTROIDS

Attribute	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Epoch date	2017-04-30 22:35:13	2017-04-30 23:32:31	2017-04-30 22:16:12	2017-04-30 22:16:07
Available satellites	19.0311	15.6724	18.1718	15.7596
GLONASS satellites	8.1365	6.8342	8.3962	6.1779
GPS satellites	10.8946	8.8382	9.7746	9.5816
SBAS satellites	0	0	0.0005	0
Number of cycle slips on $L1$ and $L2$	55.6182	59.5535	18.9372	21.3757

TABLE VIII. RESULTS, CLUSTERED INSTANCES

Cluster	Clustered Instances	Percentage
1	821,240	21
2	1,394,169	36
3	758,616	20
4	900,156	23

instance, Cluster two has 39 %.

Table XI shows a summary of all the results obtained from the different tests.

Fig. 3 shows that the heap size required to download data in serial mode is between 0 and 200 MB.

We observe from Fig. 4 for one thread, the size of the available is used almost in its entirety,

Especially for the period, 11:30 to 11:40 required 120 MGB of data.

Fig. 5 to 9 show the performance of Heap Memory for each one of the tests.

The execution with eight threads in parallel between 10% and 60% of the CPU used.

The heap size reached a maximum of 1500 MB, as shown in Fig. 5.

In the configuration of 10 threads at the beginning was 85% to achieve stability of 50%. The heap size is large at the

TABLE IX. RESULTS, CLUSTER CENTROIDS GROUPED BY DAYS

Attribute	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Epoch date	2017-05-08	2017-05-01	2017-05-21	2017-04-30
Available satellites	9,83	17.731	18	16.35
GLONASS satellites	8.33	8.37	8.56	6.27
GPS satellites	1.5	9.37	9.44	10.08
SBAS satellites	0	0	0	0
Number of slips on $L1$ and $L2$	801,016.67	496,652.85	2,431,711.03	1,396,281.27

TABLE X. RESULTS, INSTANCES GROUPED BY DAYS

Cluster	Clustered Instances	Percentage
1	6	4
2	52	39
3	27	20
4	49	37

TABLE XI. TEST RESULTS

Test	Number of threads	Average execution time (minutes)	Speedup
Sequential	1	101,15	1X
Parallel	8	48,34	2,09X
Parallel	10	46,27	2,19X
Parallel	16	67,06	1,51X
Parallel	32	474,58	0,21X
Parallel	50	522,86	0,19X

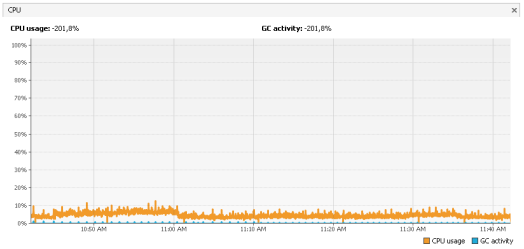


Fig. 3. CPU Performance for One Thread.

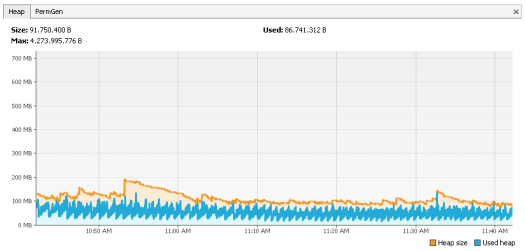


Fig. 4. Heap Memory Performance for One Thread.

Fig. 5. *a)CPU Performance- b) Heap Memory Performance*, Eight Threads.

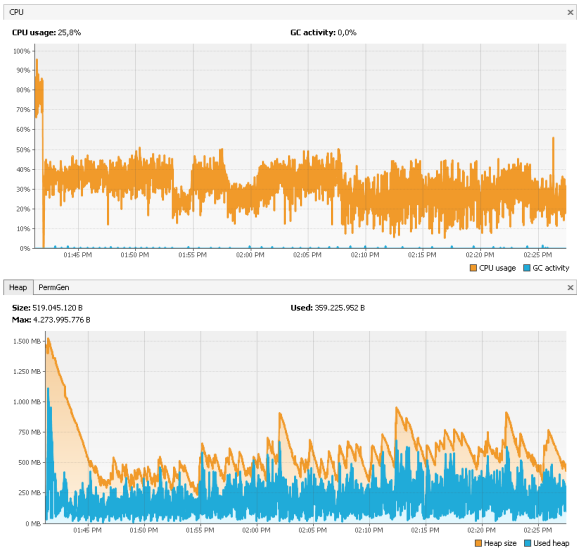


Fig. 6. *a) CPU Performance; b) Heap Memory Performance* , 10 Threads.

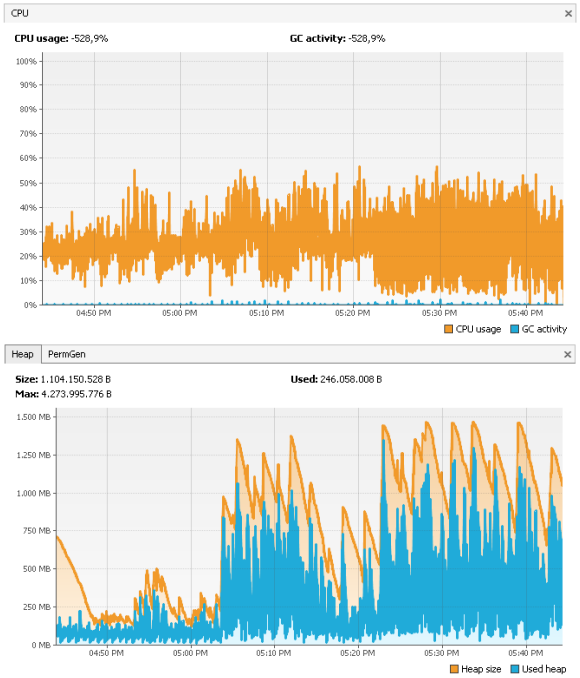


Fig. 7. a) CPU Performance, b) Heap Memory Performance, 16 Threads.

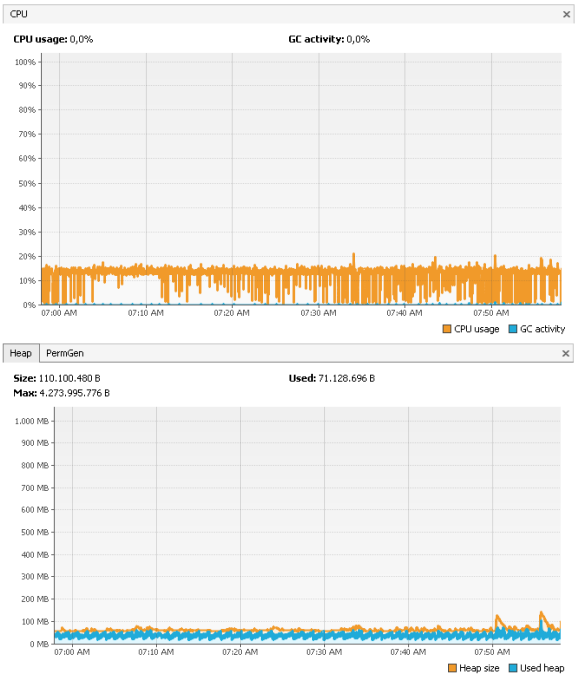


Fig. 8. *a) CPU Performance; - b) Heap Memory Performance*, 32 Threads.

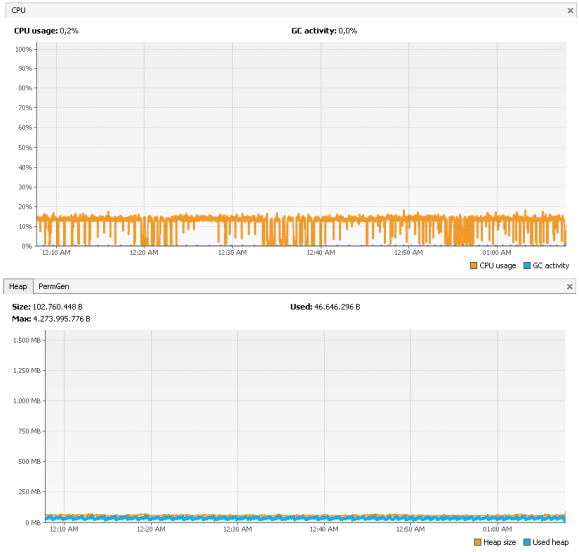


Fig. 9. *a) CPU Performance; b) Heap Memory Performance*, 50 Threads.

outset and stabilizes at around 2 pm when the heap used is lower by a difference less than 100 MB, as shown in Fig. 6.

The use of the CPU with 16 threads was 53% on average. It remained low until after 5 pm when it reached 1500 MB, giving a ratio of close to one for the heap used, as shown in Fig. 7.

Less than 20% of the CPU was for 32 threads.

A very close to 20 MB download between the heap size and the heap used, as shown in Fig. 8.

For 50 threads, the CPU usage was between 1% and 15%, and the heap size and memory used were correlated, as shown in Fig. 9.

Based on the results, the application's performance did not improve even for a large number of threads. However, an improvement was the program with eight or ten threads increased the yield by a factor of almost three. It is because parallel programs depend on finding the best balance between the available hardware and the compiler.

It was evident that when more cores used and the workstation's capacity exceeded, more time was required to execute the processes to create dynamic resource allocation queues.

The permanent space memory remained mostly unchanged throughout these tests, with peaks at 25 MB when the hardware was optimized. That is when the effective use of the multi-processors was optimized. In contrast, the heap space showed various changes in the tests. In serial mode, it showed an almost regular size of around 128 MB. In parallel mode with eight to 10 threads, it reached a peak at the beginning of the execution of a little over 1000 MB.

It was due to the operating system's initial allocation of resources as threads distributed among the cores.

An average load of about 450 MB was then required for balance, with several peaks of up to almost 700 MB due to the files' different sizes.

For 16 threads, there was a considerable amount of memory usage, including high consumption of space for three-quarters of the running time and an average of approximately 1000 MB. The process queue achieved rapid allocation of resources by the operating system. It handled many objects in the memory since memory recovery was slower than the instantiation of objects in the application. When testing more than 16 threads, constant values obtained of around 90 MB in heap space, which we interpreted as indicating that the process queue was massive. The operating system, therefore, distributed the workloads between several cores without allocating higher priorities. The outcome was that these values did not change throughout the implementation and showed lower performance than the sequential mode.

VI. Conclusions

This article proposes a method for extracting observations from several global positioning systems in real-time; these data are valuable for various science areas. Our method features a four-layer architecture. Experiments comprising a sequential test, a parallel test with multiprocessors, a clustering test, and a semantics queries test were designed and conducted. The results show a performance improvement of three-fold when running the program with eight or ten threads. Our dataset was about 100 GB in size, and retrieval was achieved in less than 60 minutes.

The speed and size of the downloaded files depend on the communications network and the availability of different broadcasters, and data extraction, therefore, has an external dependency. The development of applications using programming language optimizations gives more excellent reliability and efficiency.

It is mostly useful for applications with a high burden and a high communication level with a database.

The efficient administration of drivers and meta-data makes a difference in terms of performance.

The size of the dataset used in this study was approximately 100 GB; Big Data generally deals with much larger datasets and may also include structured or unstructured datasets. However, finding and correcting observations from several global positioning systems in real-time is a Big Data problem since it contains the four aspects of Velocity, Variety, Volume, and Veracity. where (i) velocity is the speed with which the satellites publish their results in seconds (ii) variety within existing systems for navigation and geospatial positioning, and their results are heterogeneous; (iii) volume: every second has new data, so we are talking about terabyte level; iv) veracity, in theory, is real because we are analyzing the results of satellites. RINEX files to be identified with their semantics is a challenging task.

The scalability of the method depends on the storage infrastructure for the RINEX files. The percentage of errors in the downloaded files was not more significant than 2%. Given that RINEX is an information exchange file, it complies with the conditions imposed on an exchange file. Interoperability between the various operating systems, non-redundancy of data, possibility of adding new observations excepts with a fundamental one: the great length of its files.

Initially, The method may have opted for reducing its size by choosing a binary format but at the cost of losing access to its content and availability for the user.

Nowadays, file compression programs reduce the RINEX file by a factor of three or more. For example, a file of half a day of observation, with times of 30 seconds, can occupy 1.5-2 Mb and compacted to 500-600 kb. The recording of these files has a maximum of 80 characters per line, but they contain thousands of lines.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

The data used to support the findings of this study are available from the corresponding author upon request. This work was supported by the Sciences Research Council (CONACyT) through the research project number 262756 http://navigationgnssproject.net/index.html.

References

[1] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.

[2] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.

[3] S. Li, S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein *et al.*, "Geospatial big data handling theory and methods: A review and research challenges," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 119–133, 2016.

[4] R. Tardío Olmos, A. Maté, J. Trujillo *et al.*, "An iterative methodology for defining big data analytics architectures," 2020.

[5] B. Hofmann-Wellenhof, H. Lichtenegger, and E. Wasle, *GNSS–global navigation satellite systems GPS, GLONASS, Galileo, and more*. Wien; New York: Springer, 2008. [Online]. Available: http://dx.doi.org/10.1007/978-3-211-73017-1

[6] S. Yalvac and M. Berber, "Galileo satellite data contribution to gnss solutions for short and long baselines," *Measurement*, vol. 124, pp. 173–178, 2018.

[7] C. Zhou, S. Zhong, B. Peng, J. Ou, J. Zhang, and R. Chen, "Real-time orbit determination of low earth orbit satellite based on rinex/doris 3.0 phase data and spaceborne gps data," *Advances in Space Research*, vol. 66, no. 7, pp. 1700–1712, 2020.

[8] K. Maciuk, "Gps-only, glonass-only and combined gps+ glonass absolute positioning under different sky view conditions," *Tehnički vjesnik*, vol. 25, no. 3, pp. 933–939, 2018.

[9] X. Li, M. Ge, X. Dai, X. Ren, M. Fritsche, J. Wickert, and H. Schuh, "Accuracy and reliability of multi-gnss real-time precise positioning: Gps, glonass, beidou, and galileo," *Journal of Geodesy*, vol. 89, no. 6, pp. 607–635, 2015.

[10] J. Wang, "Stochastic modeling for real-time kinematic gps/glonass positioning," *Navigation*, vol. 46, no. 4, pp. 297–305, 1999.

[11] M. Rieke, T. Foerster, J. Geipel, and T. Prinz, "High-precision positioning and real-time data processing of uav systems," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, pp. 119–124, 2011.

[12] X. Li, G. Dick, C. Lu, M. Ge, T. Nilsson, T. Ning, J. Wickert, and H. Schuh, "Multi-gnss meteorology: real-time retrieving of atmospheric water vapor from beidou, galileo, glonass, and gps observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 12, pp. 6385–6393, 2015.

[13] G. Wübbena, A. Bagge, G. Seeber, V. Böder, P. Hankemeier *et al.*, "Reducing distance dependent errors for real-time precise dgps applications by establishing reference station networks," in *Proceedings of Ion Gps*, vol. 9. Institute of Navigation, 1996, pp. 1845–1852.

[14] H. A. Karimi, D. Roongpiboonsopit, and H. Wang, "Exploring real-time geoprocessing in cloud computing: Navigation services case study," *Transactions in GIS*, vol. 15, no. 5, pp. 613–633, 2011.

[15] J. Liu, B. Priyantha, T. Hart, H. S. Ramos, A. A. Loureiro, and Q. Wang, "Energy efficient gps sensing with cloud offloading," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. ACM, 2012, pp. 85–98.

[16] D. Lyu, F. Zeng, X. Ouyang, and H. Zhang, "Real-time clock comparison and monitoring with multi-gnss precise point positioning: Gps, glonass and galileo," *Advances in Space Research*, vol. 65, no. 1, pp. 560–571, 2020.

[17] V. Kopytov, A. Shulgin, N. Demurchev, P. Kharechkin, and V. Naumenko, "High-speed stream data collection and processing system of the earth's ionospheric sounding," in *IOP Conference Series: Materials Science and Engineering*, vol. 450, no. 2. IOP Publishing, 2018, p. 022005.

[18] M. Kakooei and A. Tabatabaei, "A fast parallel gps acquisition algorithm based on hybrid gpu and multi-core cpu," *Wireless Personal Communications*, vol. 104, no. 4, pp. 1355–1366, 2019.

[19] D. Agrawal, S. Das, and A. El Abbadi, "Big data and cloud computing: current state and future opportunities," in *Proceedings of the 14th International Conference on Extending Database Technology*. ACM, 2011, pp. 530–533.

[20] C. Boja, A. Pocovnicu, and L. Batagan, "Distributed parallel architecture for" big data"," *Informatica Economica*, vol. 16, no. 2, p. 116, 2012.

[21] X.-w. DING, S.-l. XIE, and J.-y. LI, "Parallel etl based on spark," *Computer Engineering and Design*, p. 09, 2017.

[22] M. Bala, O. Boussaid, and Z. Alimazighi, "A fine-grained distribution approach for etl processes in big data environments," *Data & Knowledge Engineering*, vol. 111, pp. 114–136, 2017.

[23] X. Liu, C. Thomsen, and T. B. Pedersen, "Mapreduce-based dimensional etl made easy," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1882–1885, 2012.

[24] M. Radonić and I. Mekterović, "Etlator-a scripting etl framework," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2017, pp. 1349–1354.

[25] X. Liu, C. Thomsen, and T. B. Pedersen, "Cloudetl: scalable dimensional etl for hive," in *Proceedings of the 18th International Database Engineering & Applications Symposium*. ACM, 2014, pp. 195–206.

[26] M. Bala, O. Boussaid, and Z. Alimazighi, "P-etl: Parallel-etl based on the mapreduce paradigm," in *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*. IEEE, 2014, pp. 42–49.

[27] M. F. Masouleh, M. A. Kazemi, M. Alborzi, and A. T. Eshlaghy, "Optimization of etl process in data warehouse through a combination of parallelization and shared cache memory," *Engineering, Technology & Applied Science Research*, vol. 6, no. 6, pp. 1241–1244, 2016.

[28] C. Thomsen and T. B. Pedersen, "Easy and effective parallel programmable etl," in *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*. ACM, 2011, pp. 37–44.

[29] P. S. Diouf, A. Boly, and S. Ndiaye, "Performance of the etl processes in terms of volume and velocity in the cloud: State of the art," in *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. IEEE, 2017, pp. 1–5.

[30] J. Han, E. Haihong, G. Le, and J. Du, "Survey on nosql database," in *Pervasive computing and applications (ICPCA), 2011 6th international conference on*. IEEE, 2011, pp. 363–366.

[31] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[32] C. Ji, Y. Li, W. Qiu, U. Awada, and K. Li, "Big data processing in cloud computing environments," in *Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on*. IEEE, 2012, pp. 17–23.

[33] W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," *ACM sIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1–5, 2013.

[34] R. R. Vatsavai, A. R. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, and S. Shekhar, "Spatiotemporal data mining in the era of big spatial data: algorithms and applications," in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, BigSpatial@SIGSPATIAL 2012, Redondo Beach, CA, USA, November 6, 2012*, 2012, pp. 1–10.

[35] L. Self, "Use of data mining on satellite data bases for knowledge extraction." in *FLAIRS Conference*, 2000, p. 149–152. [Online]. Available: http://www.aaai.org/Papers/FLAIRS/2000/FLAIRS00-029.pdf

[36] D. R. Azevedo, A. M. Ambrosio, and M. Vieira, "Applying data mining for detecting anomalies in satellites." IEEE, May 2012, pp. 212–217.

[37] S.-S. Ho and A. Talukder, "Automated cyclone discovery and tracking using knowledge sharing in multiple heterogeneous satellite data," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, p. 928–936. [Online]. Available: http://dl.acm.org/citation.cfm?id=1402001

[38] Y. Li, Y. Wang, J. Yan, and Y. Qi, "The application of data mining in satellite TV broadcasting monitoring." IEEE, Apr. 2009, pp. 357–359.

[39] S. C. Tay, W. Hsu, K. H. Lim, and L. C. Yap, "Spatial data mining: Clustering of hot spots and pattern recognition," in *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International*, vol. 6, 2003, p. 3685–3687.

[40] Y. Guo, J. G. Liu, M. Ghanem, K. Mish, V. Curcin, C. Haselwimmer, D. Sotiriou, K. K. Muraleetharan, and L. Taylor, "Bridging the macro and micro: A computing intensive earthquake study using discovery net." in *SC*. IEEE Computer Society, 2005, p. 68.

Deep Reinforcement Learning based Handover Management for Millimeter Wave Communication

Michael S.Mollel[1], Shubi Kaijage[2], Michael Kisangiri[3]
The Nelson Mandela African Institution of Science and Technology (NM-AIST)

***Abstract*—The Millimeter Wave (mm-wave) band has a broad-spectrum capable of transmitting multi-gigabit per-second date-rate. However, the band suffers seriously from obstruction and high path loss, resulting in line-of-sight (LOS) and non-line-of-sight (NLOS) transmissions. All these lead to significant fluctuation in the signal received at the user end. Signal fluctuations present an unprecedented challenge in implementing the fifth generation (5G) use-cases of the mm-wave spectrum. It also increases the user's chances of changing the serving Base Station (BS) in the process, commonly known as Handover (HO). HO events become frequent for an ultra-dense dense network scenario, and HO management becomes increasingly challenging as the number of BS increases. HOs reduce network throughput, and hence the significance of mm-wave to 5G wireless system is diminished without adequate HO control. In this study, we propose a model for HO control based on the offline reinforcement learning (RL) algorithm that autonomously and smartly optimizes HO decisions taking into account prolonged user connectivity and throughput. We conclude by presenting the proposed model's performance and comparing it with the state-of-art model, rate based HO scheme. The results reveal that the proposed model decreases excess HO by 70%, thus achieving a higher throughput relative to the rates based HO scheme.**

***Keywords*—Handover management; 5G; machine learning; reinforcement learning; mm-wave communication**

I. Introduction

Unlike its predecessors, the fifth-generation (5G) of mobile communication networks has been considered a paradigm shift due to its attractive service in terms of latency, data rates, device inter-connectivity, and network flexibility. These enhancements in Key Performance Indicators (KPIs) make 5G a game-changer by allowing new applications such as remote surgery, smart cities, device-to-device communication (D2D), industrial Internet, smart agriculture, etc. [1].

To meet these service requirements and demands, 3GPP has launched the New Radio (NR) standardization with the following use cases: enhanced mobile broadband (eMBB), massive Machine Type Communication (mMTC), and Ultra-Reliable Low-Latency Communication (URLLC) [2], [3]. eMBB aims at enhancing the system capacity and supporting the ever-increasing end-user data rate. eMBB introduces two significant technological enhancements: mm-wave use to achieve higher data rate and antenna array that supports massive multiple-input and multiple-output (MIMO) beamforming. URLLC introduces entirely new use-cases requirements to support vertical industries such as self-driving cars, remotely surgery for eHealth and other mission-critical use cases. The unique features introduced by URLLC include improved latency, reliability while guaranteeing high service availability and security. mMTC intends to provide cost-efficient and robust connection of billions of devices that transmit small packets of data (with 10s latency) but without overloading the network. Some factor to consider in mMTC are low power consumption, longtime availability of service, and coverage. mMTC can also be seen as a particular case of URLLC with more emphasis placed on reliability while less emphasis is placed on the latency [3]. The new use cases pave the way for increasing interconnected devices to the Internet, resulting in the Intenet of Things (IoT) development. IoT is a technology that targets to connect everyday devices (e.g., home appliances, wearable devices) to the Internet, making the scenario even severer. The considerable projection increase in the number of cellular IoT devices in the near future [4] entails 5G networks dealing with stringent requirements and an increasing number of connected devices.

Heterogeneous network (HetNet), Ultra-Dense Network (UDN) and the use of mm-wave are candidate solutions to overcome the possible challenges of 5G networks [5]. Together, they can significantly increase network throughput, available bandwidth and spectral efficiency [6]. HetNet is the deployment of various base station (BS) topology based on coverage footprint and type of Radio Access technology used [7]. Moreover, densification of the network is a phenomenon of deploying more small cells (SCs) in the network to increase cell density, coverage, and network throughput. The main challenge of deploying UDN is increased interference sources and signal fluctuation. For example, there are many access points (AP) and cells in crowded substations or stadiums; thus, signals can have more reflecting and scattering paths, contributing to high signal interference and fluctuation. On the other hand, the concept of utilizing a broader bandwidth refers to opening up a new frequency spectrum for mobile communication to increase the available bandwidth. mm-Wave frequencies offer great potentials in terms of data rate due to their larger bandwidth, and mm-wave bands have been designated as Frequency Range-2 (FR2) in 5G New Radio (NR) [8]. Nevertheless, the mm-wave spectrum comes with its limitation as it is more likely to suffer from extreme penetration losses due to higher frequencies. Thus, mm-wave use as carrier frequency decreases the BS footprint area, thereby resulting in multiple SCs in the network.

Network densification is an inevitable destination for network operators to provide a more sustainable and enhanced Quality of Service (QoS) for mobile users. However, network densification with SCs is not a solution without any side effects; it increases the number of HOs, which is characterized by changing from one BS to another BS for the user equipment (UE) when there is an ongoing communication (voice or data). Given the limited coverage area of SCs, the UE would

need more HOs since there will be more BSs within an area of interest after the densification. Moreover, the different types of BSs from HetNet deployment will result in complicated HO signalling processes [6]. Furthermore, considering that the HO interval is inversely proportional to the UE speed [9], the case becomes even more severe in the case of high mobility user.

HO process involves exchanging information between serving BS, target BS and Core Network (CN). Exchanged messages, commonly known as signalling overheads, is necessary during the three (3) steps involved in HO, which are HO preparation, execution and completion. If excessive and undesirable HO increases, then both signalling overhead and average HO interruption time increases [10], [11]. The high signalling overhead and HO interruption time result in a significant increase in latency, thereby undermining the attempt to meet with 5G network specifications, particularly the URLLC use cases. Besides, the average throughput also decays with the increasing number of HOs, resulting in degraded quality of experience (QoE) for the users [12]. Therefore, it is apparent that special consideration should be given to HO management to ultimately achieve and unleash the potential of the 5G networks by meeting all its requirements.

To meet the 5G expectation, novel and advanced HO control that minimizes the effects of HO are required. The focus is on reducing unnecessary, and unwanted HO events such as ping-pong and frequent HOs, and the main parameters to be considered are the total number of HOs per UE trajectory and the time spent during HO. These parameters together define HO cost, which is the multiplication of both parameters [12]. In other words, the total number of HOs and the time spent during a single HO should be reduced to get away with one of the negative implications of using mm-wave spectrum in the UDNs. The former can be achieved through an intelligent method by avoiding 'unnecessary' HOs, whereas the latter is a characteristic of the RAT [13]. Therefore, in this paper, we present an intelligent method based on DRL for HO reduction in mm-wave BSs in a UDN environment.

The rest of this article is organized as follows. First, in Section II, we describe HO management in 5G networks, and a review of the state-of-the-art HO management approaches, then in Section III, the Deep Reinforcement Learning (DRL) framework was introduced as well as how it is linked to HO problem. Next, in Section IV, the use case is presented as well as a description of the simulation environment. In Section V, we evaluate the performance of the proposed model and compare it with the rate based HO scheme. Finally, Section VI concludes the paper.

II. HO Management in 5G Networks

HO is described as the process of transferring an ongoing UE's resource from one channel to another in wireless mobile communication. The process mainly involves a change of connection from either serving BS, carrier frequency channel or prioritizing a new technology found within the UEs' vicinity. One of the key design strategies for the successful implementation of 5G networks is the efficient handling of HO to make UEs seamlessly change BS association, thereby limiting unnecessary HO. HO process in mobile communication involves three states. The first stage is the measurement or information gathering phase, where the UE measures the signal strength (other parameter measurements are also possible) of every potential neighbour BS and the current serving BS. The second phase is about the HO decision, where the current serving BS decides to initialize the HO based on the measured data from the first stage. The third phase is the cell exchange phase, when the UE releases the serving BS and connects to the new BS [14].

Traditionally, HO is of two types, hard and soft HOs. In the case of hard HO, the connection must be released from the serving BS before the connection with the target BS can be established. In soft HO, the serving BS connection is maintained and used for a while in parallel with the target BS connection [14]. 5G mm-wave communication supports the hard HO method in most cases [8]. Besides, it supports dual connectivity, which means that the UE can be connected to more than one BS. However, when it comes to HO in dual connectivity, the individual connections perform hard HO, and new HO scenarios emerge, which lead to more HO complications in mm-wave communication [15].

Mm-wave communication is already severely affected by blockages and high path loss; thus, deploying multiple mm-wave BSs would result in additional challenges, particularly from HO management's perspective. Hence, by adopting hard HO in mm-wave communication, the UE will often experience intermittent connections, leading to poor QoE regardless of QoS. One of the causes of UE dissatisfaction from mm-wave BS might be either blockage or interference, leading to a reduction in the SNR of the serving BS; these situations present a ping-pong problem. Another cause of UE dissatisfaction from mm-wave BS is when UE moves out of signal range since it is known that the UE experiences excellent coverage of mm-wave communication when it is within 200m from the serving BS [16]. The challenge is selecting BS intelligently during HO in such a way that leads to a few ping pong, reducing unnecessary HO, and maintaining UE-BS connectivity for a long duration. Generally, optimal BS selection to re-associate with UE is needed to reduce the problem mentioned above. In legacy technology, fourth generation and all technology which use sub 6 GHz, the issue of HO is less severe considering the sparse nature of BS deployment compared to 5G, which uses mm-wave frequencies. Furthermore, sub 6 GHz has a broad coverage compared to mm-wave, making unnecessary HO less frequent. It is worth noting that the HO process involves several procedures, but we present the general conditions required for HO to occur for the sake of simplicity.

A. HO Process in 5G

In 5G, 3GPP [8] defines six HO events for entering and leaving. These events are A1, A2, A3, A4, A5, and A6 and are used to trigger HO. They are described as follows [17]: Event A2 and A1 are activated when the UE's channel condition drops below and exceeds the configured threshold, respectively. They are also used to start and stop inter-frequency neighbour search. Intra-frequency HO is initiated by event A3 when the neighbouring channel's condition is higher than the service channel's condition based on the configured threshold. Event A4 and A5 are typically used for inter-frequency HO, where the target cell's signal strength has to be higher than the absolute threshold for the A4 event to be triggered. In addition to

Event A4, however, event A5 requires that the serving BS radio frequency (RF) condition be below a certain threshold. Event A6 is similar to event A3 but is used for intra-frequency HO to the secondary frequency on which the UE is encamped. Event A4 and A5 can also be used for conditional HO management, e.g. load balancing. Event B1 and B2 specifies the entering and leaving condition for inter-RAT HO [8]. The threshold values are all configured value, and if they are correctly configured, they can significantly reduce the number of unnecessary HOs. In this paper, we assume for UE to HO, one of the trigger conditions for HO must be met.

However, these HO events only show the minimum requirements for the UE to undergo HO. The HO trigger events do not include any intelligence in deciding which BS to associate UEs with, especially when choosing among multiple BSs. Hence, it always chooses the BS that provides the highest empirical rewards, for instance, BS with the highest signal to interference plus noise ratio (SINR) or highest reference signal received power (RSRP). Furthermore, the selection of the optimal BS to HO does not only depend on the BS which provides the maximum instantaneous reward SINR but other factors such as throughput, which depends on bandwidth and number of UEs, also need to be considered, especially for mm-wave BS, thus, making the matter of optimal BS selection an open issue.

The conventional event-based HO trigger depends only on the UE's measurement report (MR) rather than the general network perception, which often results in sub-optimal HO decisions. Moreover, in 5G, HO decisions would be taken at the network level, where both the distribution and load of users alongside BSs status would be considered. Intelligence is therefore required to make optimum decisions regarding selecting the target BS by incorporating or considering other appropriate features during the BS selection process.

B. State-of-the-Art HO Management Approaches

In [18], the authors addressed the HO prediction method in 5G and used RL to find the optimal beam that the UE should select to maximize throughput. Their method assumes that the state fed to RL is the combination of all RSRP values seen from all surrounding BSs. However, considering the states as discrete values in such a complex environment, the proposed solution does not generalize the HO solution. The states created by combining RSRP are continuous intrinsic values and not discrete values as assumed. The actual network generates continuous RSRP values.

More recently, there have been several studies that solve HO using multi-armed bandit. The armed bandit is the classic probability-based RL problem. In [19] the authors assume the UE as an agent and set the BS as an arm which the UE chooses to maximize its return, which is the average throughput for their case. The dynamics of the environment was well-considered and captured in the learning process. However, they only considered UE dynamics in their work without considering the dynamics of the environment, such as moving and stationary obstacles, which can make the solution more complex. They also did not consider user trajectory. Despite the success of [20] in optimizing HO from an energy point of view, the proposed model is still insufficient as it ignores some vital factors such as UEs trajectory and distribution as well as the available bandwidth in the target BS.

In addition, different heuristic approaches have also been proposed as an alternative solution to the HO problem. Several researchers have focused their attention on different HO management techniques using these approaches. For example, [14] demonstrates how inter-cell interference coordination (ICIC) can be used to enhance HO decision performance. There is also a more advanced version of ICIC known as enhanced Inter-Cell Interference Coordination (eICIC), which can reduce the HO failure ratio (HoF) and the radio link failure (RLF) compared to the case without eICIC. However, despite the advantages of this method, it involves extensive overhead signalling during coordination between the BS and finding the global solution regarding when and which BS to HO, thereby increasing delay and degrading UE's QoE. A BS skipping technique for mobile UEs that demonstrates a significant increase in the overall UE throughput was proposed in [12]. The authors take advantage of a coordinating BS in deciding which BS to select to reduce the number of HOs. They also added a HO cost function, which penalizes the action of HO and maintains the minimum SINR as much as possible to avoid taking HO. Their method has been proven to work based on stochastic analysis, but the fundamental question remains how to skip BSs smartly. Hence, there is a need to develop intelligent BS skipping techniques which incorporates all the necessary factors during decision making.

In order to overcome the stated challenges while achieving high throughput in mm-wave communication, we propose a DRL algorithm that intelligently selects the BS that will prolong UE-BS association while guaranteeing maximum throughput. We develop an efficient method that alleviates the effect of HO and help realize the potential of mm-wave frequency in 5G systems. We leverage the availability of extensive data that the network generates during the training phase. The advantage of the proposed method is that it learns offline before its deployment to the BS controller to assist in HO decision. The model aims to maximize the system's average throughput by considering the signal to noise ratio (SNR), UE velocity, number of HOs per UE trajectory, and network load balancing.

III. Reinforcement Learning Assisted HO Management

Our objective is to achieve the maximum throughput, which is achieved if the whole network environment is considered. The network environment includes, but is not limited to, UE trajectories, velocity and distribution, blockages, and BS distribution and UE velocity. Some of these factors vary with time, while others do not. Therefore, it is difficult for the heuristic approaches to solve the HO problem while including changing factors over time. Hence, the solution is to explore the environment and exploit the actions that achieve the intended objective. Artificial Intelligence (AI) has a class of algorithms known as RL that solves this problem; these algorithms learn through trial-and-error. When combined with Deep Neural Network (DNN), RL forms DRL, which performs exhaustive search and learns by themselves through experience from interacting with the environment to achieve the objective of maximizing or minimizing the objective function.

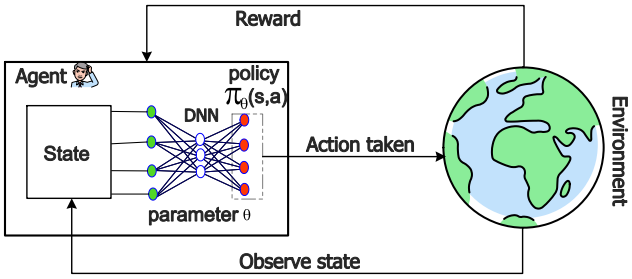


Fig. 1. General Framework of RL.

A. Reinforcement Learning

This section gives a brief overview of the RL and DRL framework and further discusses how the HO optimization problem is formulated and solved using the DRL algorithm.

1) RL Framework: RL is a subfield of AI that enables machines to create artificially intelligent agents that learn to optimize their accumulated reward by interacting with the environment. In RL, the agent receives feedback after each action. The feedback includes the reward and the next state of the environment. The relationship between agent, action and environment is shown in Fig. 1 [21]. The agent learns the best policy through multiple interactions with the environment, and the learning procedure is detailed in the following paragraphs.

Here, we first define the main elements of RL. At time t, the agent observes the state of the environment, $s_t \in S$, where S is the set of possible states. After observing state s_t agent takes an action, $a_t \in A(s_t)$ where $A(s_t)$ is the set of possible actions at state s_t. After selecting and taking the action a_t from state s_t, agent receives the immediate reward r_{t+1} from state-action pair (s_t, a_t). The selected action in state s_t moves the agent to state s_{t+1} at time $t + 1$. It is essential for the environment to have state dynamics such that $P(s_{t+1}|s_t, a_t)$ exists. There are two approaches to solving RL problems: The first approach is based on policy search, and the second approach is based on the value function approximation. Their names reflect their behaviour. The former searches directly for the optimal policy based on a parameterizing policy such as NN. The later keeps improving the value function estimate by selecting actions greedily according to the previously updated value function and indirectly learning optimal policy.

RL methods have a dilemma, which is the trade-off between exploitation and exploration. This has to do with how the agent learns the environment through trial and error. Should the agent be encouraged to perform exploitation or exploration during learning? Exploitation implies that the agent acts more greedily by taking the best actions that maximize the reward. Exploration means the agent act less greedy, so it can learn about the environment more to find optimal actions. The most common solution to this dilemma is the e-greedy policy where the agent explores with probability less than $\epsilon \in [0, 1]$ and exploits the best action otherwise is applicable to value function. For policy search methods, the problem is less severe.

2) DRL Framework: All RL methods based on tabular solution suffer from the so-called "the curse of dimensionality", which means that computational requirements increase exponentially with an increase in the number of states. Moreover, for the task involving continuous states, the problem becomes severe. To overcome this problem, DRL is introduced by exploiting the advantage of neural networks (NN) in the traditional RL. The idea behind DRL is to train neural networks to approximate optimal policy [21].

In [22], the authors combine deep convolutional neural networks (CNN) with RL to develop a novel artificial agent capable of learning successful policies directly from high-dimensional sensory input data. The CNN is used to represent the action-value function, denoted as $Q(s, a; \theta)$, where $Q(s, a)$ represents the action-value function and the parameter θ is the weight of the neural networks. θ is updated every time Q - network performs an iteration with the mean square error as the loss function. The loss function is the mean square error between the action-value $Q(s, a; \theta)$ and target values $r + \gamma \cdot \max_{a'} Q^*(s', a'; \theta^-)$.

It is imperative to train the neural network using training samples from both the previous and current episodes. This is necessary because approximating the optimal policy direct using only current samples results in slower learning and undesirable temporal correlations. To solve this problem, the concept of experience replay, in which previous experiences by the agent at each time-step (s_t, a_t, r_t, s_{t+1}) as well as recent experience are stored for subsequent use in the training phase. The experience replay buffers previous experiences and randomly selects the training set over the data. This results in the gradual smoothing of the data distribution to avoid the bias of the sample data.

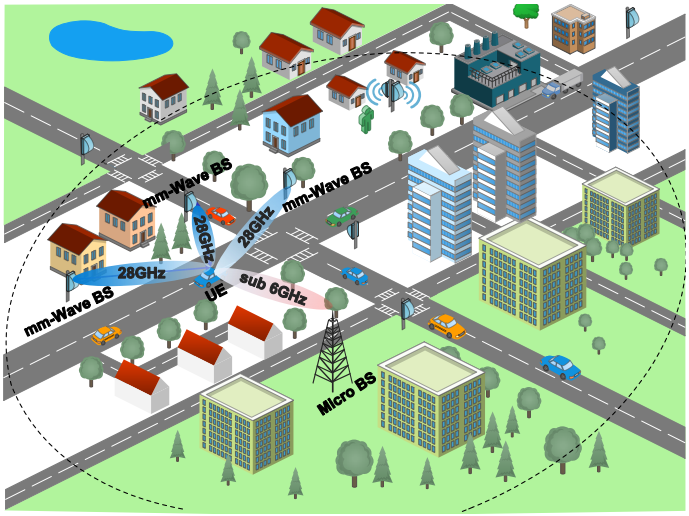


Fig. 2. Overview of Heterogeneous Network (HetNets) with Dense mm-wave BS, UE's and sub 6 GHz BS in the Urban Area.

IV. DRL-Aided Intelligent BS Selection

In this section, we explain our considered system model. Then, we describe the proposed DRL optimal BS selection framework. It is worth noting that the DRL framework is based on Deep Q Network (DQN) and that both terms would be used interchangeably for the rest of this paper.

A. System Model

We consider Fig. 2 as our use-case system model, which demonstrate a simplified 5G HetNet where the mm-wave SCs are placed close to each other as part of the HetNet. For simplicity, we assume that every BS and UE has a single antenna and 28 GHz, 2.1 GHz are used for mm-wave BS and sub-6 GHz BSs respectively. The environment consists of a sub-6 GHz macro BS, UE's, and the mm-wave BSs in Fig. 2. Wireless Insite (WI) software is used to develop the environment, and it uses ray tracing, which provides accurate results that mimic the actual network environment. SINR is a popular metric for measuring channel quality. In the system model, however, we consider SNR, and the reason is that mm-wave antennas are capable of forming directional beams; therefore, Inter-cell interference contribution is assumed to be negligible.

B. Proposed Optimal Base Station Selection based on DRL

In this section, we present our design and the proposed DRL-based architecture. Fig. 3 shows the main components of the proposed DRL framework, and the description of each component is presented in the following session.

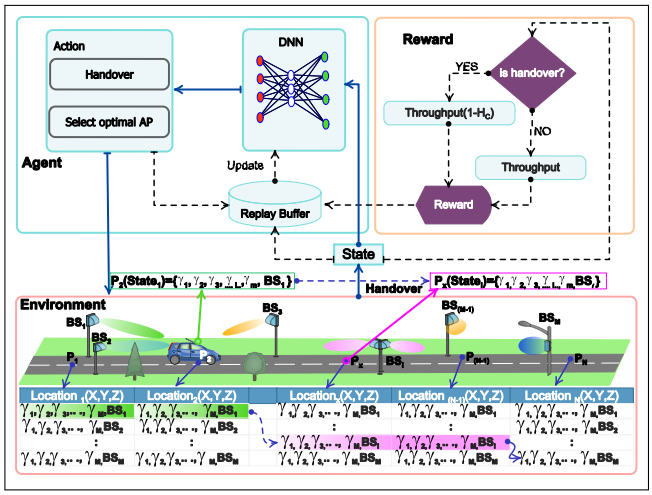


Fig. 3. DRL-based Framework Comprising Environment, States, Actions, and Rewards.

a) Agent: An agent is an entity that can interact with the environment. It observes the state of the environment, takes action and receives the consequence of the action taken. For this problem, we model the agent as a BS controller, and the reason for doing this is because the DRL model requires training resources. The BS controller is chosen because it possesses resource in terms of time, computation power, data set, and, more crucially, the entire network's global information consisting of mm-wave BSs. It should also be noted that the UE collects the input state features in the measurement report (MR) and shares them with the agent.

b) Action: In the HetNet, the association strategy between UE and BS mainly depend on the HO events A1-A6 [23]. However, always choosing the target BS with the highest SNR or RSRP lead to the sub-optimal decision. The wireless environment's dynamic nature is correlated with mobile and stationary obstacles, the presence of several nearby mm-wave BS, and signal fluctuation due to path loss. These factors increase the number of HOs for mobile UE unless appropriately handled. Fig. 3 shows M mm-wave BS, and arbitrary UEs, moving from point P_1 to P_N, and in each point, P_x(X, Y, Z) is in cartesian coordinates. Intuitively, there are more than one BSs that if the UE connects to it, it can prolong UE connectivity with fewer HOs and guarantee maximum user throughput. Hence, we define the action $a \in A(s)$ as the scalar representation of the serving BS at state s. The action space $A(s)$ includes all BSs along the UE route.

c) State space: The state explains the current condition of the network environment and determines what happens next. For our problem, the state is the UE Cartesian coordinate point P_x. However, due to the difficulties involved in localizing mobility location, SNR is chosen instead to represent Point $P_x(X, Y, Z)$. We consider SNR received from all BSs at Point P_x to represent location P_x instead of actual P_x in Cartesian coordinates. Logically, the combination of SNRs from BSs is unique continuous values that are the same as point P_x in the Cartesian coordinates throughout the UE route. Therefore, we can relate UE's current position to a combination of BSs SNR values. The advantage of SNR is that UE always receives MR containing accurate SNR from the serving and neighbouring BSs, and we can use this potential information.

Hence, at point P_x, the state space for an arbitrary UE is given as, $s = \{\gamma_1, \gamma_2, \gamma_3 ..... \gamma_m, BS_{i \in m}\}$ where γ_i is the SNR of BS i, i is the index variable in m BS, and $BS_{i \in m}$ is a serving BS index in one-hot encoded vector. One-hot encoding [24] is the vector transformation of an integer variable into the binary value of zeros except for the index of the integer. For instance, if the serving Bs index at point P_x is $BS_{i=3}$ and there are a total of five BSs $m = 5$, hence, it's equivalent one-hot encoding vector become $BS_{(i=3)} = [0, 0, 1, 0, 0]$.

d) Reward Design: The reward is an abstract term reflecting environmental feedback. The importance of reward is to motivate the agent to learn to reach the target through reward maximization, and our goal is to maximize UE throughput while minimizing HOs. It is also essential to design the reward in such a way that it avoids giving delayed rewards since it may cause the so-called credit assignment problem [20], [21]. We introduce an immediate reward function estimating the immediate impact of the action taken to achieve the agent's target. We design the immediate reward so that the number of HOs and instantaneous received SNR value are combined. We derive the reward from the throughput equation as follows: The instantaneous throughput can be expressed as:

$$\mathbb{T} = \frac{B}{N} \times \log_2(1 + \mathcal{SNR}_i) \tag{1}$$

where B is the maximum bandwidth allocation per serving BS, N is the total number of UEs connected to the BS, and $\mathcal{SNR}_i$ is received SNR from serving BS_i. The reward is obtained by incorporating the impact of HO cost to eqn. 1. Hence, the reward can be expressed as:

$$r\left(s_{t+1}, a, s_t\right) = \begin{cases} \mathbb{T}\left(1 - \mathcal{H}_c\right), & \text{if HO occurs} \\ \mathbb{T}, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{H}_c$ is the HO cost [25] which is a unit-less quantity that is used to measure the fraction of time without useful transmission of data along the user's trajectory due to the transfer of HO signalling and the switching of radio links between serving and target BSs.

For model to work, we assume that the average SNR represents the long term experienced SNR at a particular point and that the agent uses these accurately collected SNR values to calculate the reward. We also assume the time delay values of 2 sec per HO for UE's HO from mm-wave BSs to mm-Wave BSs and 0.7 sec per HO for UE's HO mm-wave BSs to sub-6GHz BSs and vice versa [12].

1) Learning algorithm: Fig. 3 shows the proposed model framework built DQN algorithm, summarized in Algorithms 1. In this Algorithm 1, the first thing the agent does is to observe the type of service and if the SNR received from the serving BS is greater than the threshold then it maintains the serving BS else agent decides by taking action a following the ϵ-greedy policy. For a moving UE in particular , at position p, the UE takes action a according to the stated policy $\pi_\theta(s, a)$. Then, after one step of UE $p+1$, the environment generates the next state s_{P+1}. The experienced transition (s,a,r) is stored in the replay memory $\mathbb{D}$, after which the UE receives the next state (s_{p+1}) and perform action a_{p+1} determine by π_θ, and process continue until it reaches terminal state.

Algorithm 1: Deep Q-Learning

Initialize replay memory $\mathbb{D}$ to capacity $\mathbb{N}$;
Initialize action-value function Q with random weight θ;
Initialize the target action-value function $\hat{Q}$ with weight $\theta^- = \theta$
Initialize the target Q-network replacement frequency f_u;
Repeat:
Get Initial state
Assign terminal state ← False
Repeat The agent observes the state:
if *SNR of Serving BS$_s$ ≥ minimum SNR for service $\mathcal{C}_i$* **then**
  Action: ← Index of serving BS$_s$;
else
  Action: ← agent takes an action following ϵ-greedy policy;
end
The agent observe new state s_{p+1} after UE move from point p to another point $p+1$
From action $a(p)$ taken above, calculates the immediate reward $r(s(p), action(p))$ in position p
The agent stores all new experiences $(s(p), a(p), r(p), s(p+1), terminalstate)$ into the replay memory $\mathbb{D}$
Agent run experience replay once every f_u steps;
Sample random mini-batch of $\mathbb{Z}$ experience $(s(p), a(p), r(p), s(p+1),$ terminal state) from the reply memory $\mathbb{D}$;

$$\text{set } y_s = \begin{cases} r_{s(p)}, & \text{for terminal } s(p+1) \\ r_{s(p)} + \gamma \ max_{a'} \ Q(s, a'; \theta), & \text{otherwise} \end{cases}$$

Agent performs a gradient descent step on $(y_j - Q(s(p), a(p); \theta))^2$
The agent updates the DQN wight θ once every $\mathbb{C}$;
Every $\mathbb{C}$ step reset $\hat{Q} = Q$, i.e $\theta^- = \theta$;

V. Performance Evaluation

This section evaluates the proposed DRL-based algorithm's performance, but first, we describe the simulation set-up and parameters and then presenting the simulation results and discussions. We also compare the performance of the proposed DRL model and with the benchmark HO policy [23], which is rate based HO (RBH) strategy.

A. Simulation Setups

The environment, agent and reward are constructed as follows: The environment is constructed using ray tracing simulator WI, and states that are obtained from the environment consist of different number of BSs ranging from 10 - 70 BSs, random obstacle, the random walking model for UE with speed 1 - 10 ms^{-1} and UE's trajectories is of length 500 m length. Python with Keras library and TensorFlow framework was used to implement the agent, and reward is generated based on throughput as expressed in Eqn 2. The summary of the simulation parameters is presented in Table I. In addition, the hyper-parameters used in the implementation of the DQN are shown in the Table. II.

B. Results

The user's velocity was set to 8 ms^{-1}, and 10 mm-wave BSs were considered in the first experiment. Also, the SNR threshold values considered is within the range of 1 dB and 7 dB. We analyse the relationship between the number of HOs and the threshold SNR, which is the UE triggering condition to HO. Fig. 4 shows the different values of the minimum SNR against the number of HO. From the figure, it can be clearly observed that the proposed model outperforms the RBH. The minimum HO reduction gain is seen when the threshold SNR is 7. The trend shows that for any SNR, the proposed DQN based model outperforms RBH. Overall, the proposed DQN model resulted in a 70% HO reduction compared to the benchmark RBH method.

For the second experiment, we evaluate the running time for the two methods, as shown in Fig. 5. The parameters in this experiment are as follows: UE velocity = 8 ms^{-1}, and γ_{th} = 20 dB. Fig. 5 shows that all the policies follow a similar trend. It can be observed that our proposed model takes a longer time than RBH to decide the BS to HO the UE. This is because the proposed model considers more parameters when making a HO decision than the RBH method. Moreover, there is a linear relationship between increasing the number of mm-wave BS and running time for both policies.

Finally, we evaluate the proposed model's performance in terms of the number of HOs and throughput at different UE velocities in the last experiment. The experimental parameters are set as follows: γ_{th} = 20 dB, λ = 50 BSkm^{-2}, and UE velocity = 8 ms^{-1}. The average system throughput and the number of HOs for both HO management policies against the

TABLE I. SIMULATION PARAMETERS

Parameter	Value
BS intensity	10 - 70 (BS/km^2)
mm-wave frequency	28 GHz
mm-wave bandwidth	1 GHz
BS transmit power	30 dBm
Thermal noise density	-174 dBm/Hz
Delay without data transmission	0.75, 2 sec

TABLE II. DESIGN PARAMETERS FOR THE DEVELOPED DQN MODEL

Parameter	Value
Hidden layers, Neuron size	6, {32, 64, 128, 256, 64}
Activation function hidden layers	relu
Activation function output layer	linear
Initial exploration training	1
Final exploration training	0.2
Learning rate, α and Discount Factor, γ	0.01 , 0.9
Mini-batch size $\mathbb{C}$, Optimizer	32, Adam
Replay memory size, $\mathbb{D}$	10000

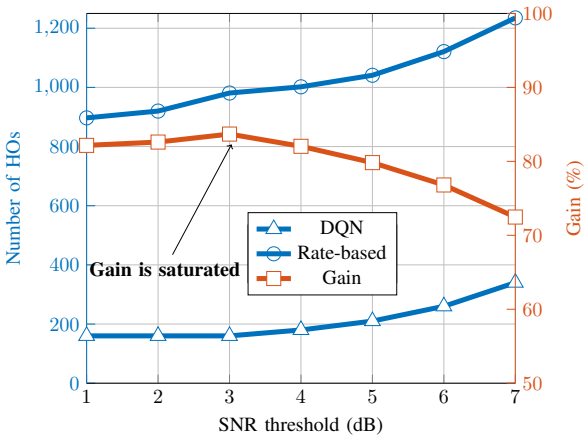


Fig. 4. Number of HO Against Different SNR Threshold.

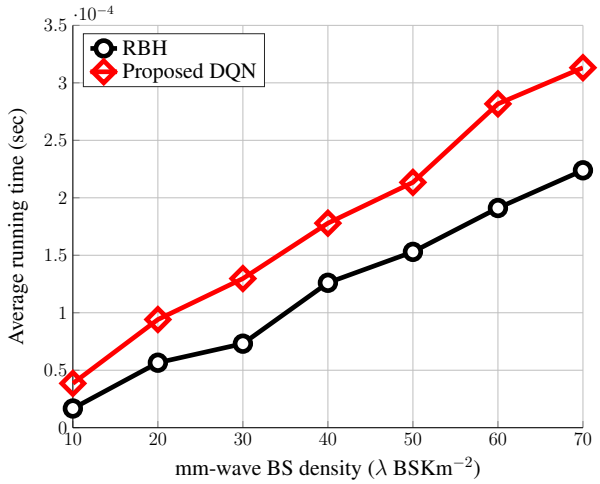


Fig. 5. Average Running Time as a Function of Number of mm-wave BS.

UE velocity are shown in Fig. 6. Fig. 6(a) shows a slight and gradual increase in the number of HOs for both models; however, the proposed DQN model outperforms the RBH policy. Compared to low-speed UE, the effect of HO on the average throughput is more significant for high-speed UE, as seen in Fig. 6(b). Nevertheless, in comparison to RBH, our model proposed performs better.

VI. CONCLUSION

Mm-wave BS deployment will become ever denser with the emergence of new 5G use cases that demand high data rate. Using mm-wave for communication between UE and BS leads to more HOs for arbitrary UE, and deploying dense mm-wave BSs increases the problem. This paper presents a DQN based model that smartly learn how to maximum UE throughput while minimizing HO's effect. The proposed DQN model and the benchmark rate based HO mechanisms are simulated, and their comparative performance analysis has been performed based on throughput and the number of HOs. According to the simulation results, it can be clearly seen that the proposed approach gives more successful results than the traditional approach in terms of throughput and number of HO occurrences. A new HO strategy that can learn by feeding various state features such as images will be presented in the future. Moreover, the idea of sharing the learnt strategy with the UEs in the learning phase in order to fasten the training process will be considered in the ultra-dense 5G network environment.

ACKNOWLEDGMENT

The authors would like to thank the African development Bank (AfDB) for funding this work.

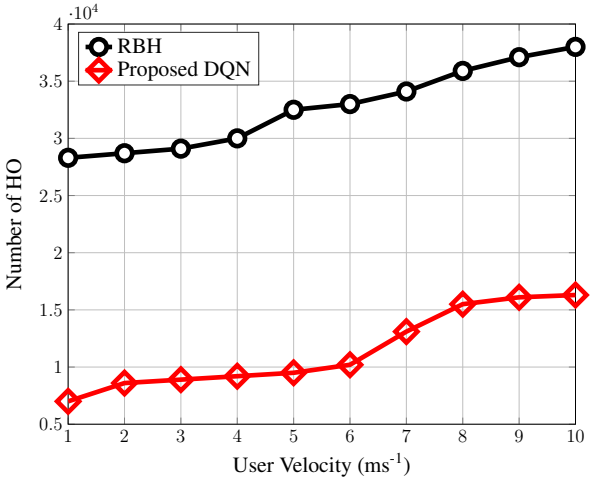
REFERENCES

[1] A. Al-Dulaimi, X. Wang, and C. L. I, *5G Communication System: A Network Operator Perspective*. Wiley, 2018, pp. 625–652.

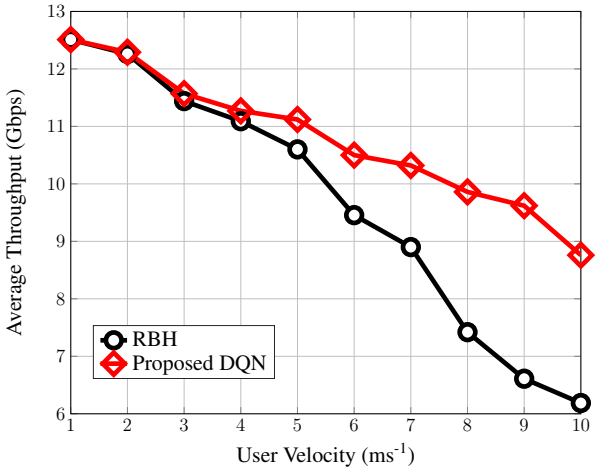
[2] 3GPP, "5G; Study on scenarios and requirements for next generation access technologies," 3rd Generation Partnership Project (3GPP), TS 38.913, Sept. 2018.

[3] S. Lien, S. Hung, D. Deng, and Y. J. Wang, "Efficient ultra-reliable and low latency communications and massive machine-type communications in 5g new radio," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–7.

[4] Ericsson, "Ericsson mobility report," Ericsson, Tech. Rep., Nov. 2018. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-november-2018.pdf



(a) The number of HOs



(b) Average system throughput

Fig. 6. Relationship between HO Performance and UE Velocity.

[5] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[6] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5g: challenges, methodologies, and directions," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 78–85, 2016.

[7] S. Dastoor, U. Dalal, and J. Sarvaiya, "Issues, solutions and radio network optimization for the next generation heterogeneous cellular network—a review," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017, pp. 1388–1393.

[8] 3GPP, "5G; NR; Base Station (BS) radio transmission and reception," 3rd Generation Partnership Project (3GPP), TS 38.104, July 2018.

[9] A. Talukdar, M. Cudak, and A. Ghosh, "Handoff rates for millimeter-wave 5G systems," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*. IEEE, 2014, pp. 1–5.

[10] E. Ndashimye, N. Sarkar, and S. Ray, "A network selection method for handover in vehicle-to-infrastructure communications in multi-tier networks," *Wireless Networks*, vol. 26, 08 2018.

[11] M. Tayyab, X. Gelabert, and R. Jäntti, "A survey on handover management: From lte to nr," *IEEE Access*, vol. 7, pp. 118 907–118 930, 2019.

[12] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Handover management in dense cellular networks: A stochastic geometry approach," in *2016 ieee international conference on communications (icc)*. IEEE, 2016, pp. 1–7.

[13] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From lte to 5g for connected mobility," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 156–162, 2017.

[14] G. Gódor, Z. Jakó, Ádám Knapp, and S. Imre, "A survey of handover management in lte-based multi-tier femtocell networks: Requirements, challenges and solutions," *Computer Networks*, vol. 76, pp. 17 – 41, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128614003715

[15] I. Shayea, M. Ergen, M. Hadri Azmi, S. Aldirmaz Çolak, R. Nordin, and Y. I. Daradkeh, "Key challenges, drivers and solutions for mobility management in 5g networks: A survey," *IEEE Access*, vol. 8, pp. 172 534–172 552, 2020.

[16] M. Attiah, M. Isa, Z. Zakaria, M. Abdulhameed, M. Mohsen, and I. Ali, "A survey of mmwave user association mechanisms and spectrum sharing approaches: an overview, open issues and challenges, future research trends," *Wireless Networks*, vol. 26, 05 2020.

[17] S. M. A. Zaidi, M. Manalastas, H. Farooq, and A. Imran, "Mobility management in emerging ultra-dense cellular networks: A survey, outlook, and future research directions," *IEEE Access*, vol. 8, p. 183505–183533, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.3027258

[18] M. Bonneau, "Reinforcement learning for 5G handover," 2017.

[19] Y. Sun, G. Feng, S. Qin, Y. Liang, and T. P. Yum, "The smart handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1456–1468, June 2018.

[20] Z. Wang, L. Li, Y. Xu, H. Tian, and S. Cui, "Handover control in wireless systems via asynchronous multiuser deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4296–4307, 2018.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[23] 3GPP, "5G;NR;Radio Resource Control (RRC);Protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.331, 10 2018, version 15.3.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197

[24] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[25] R. Arshad, H. Elsawy, S. Sorour, T. Y. Al-Naffouri, and M. Alouini, "Handover management in 5g and beyond: A topology aware skipping approach," *IEEE Access*, vol. 4, pp. 9073–9081, 2016.

Disposable Virtual Machines and Challenges to Digital Forensics Investigation

Mohammed Yousuf Uddin,[1], Sultan Ahmad*[2], Mohammad Mazhar Afzal[3]
Department of Computer Science and Engineering, Glocal University,
Saharanpur, Uttar Pradesh, India[1,3]
Department of Computer Science, College of Computer Engineering and Sciences,
Prince Sattam Bin Abdulaziz University,
Al-Kharj 11942, Saudi Arabia[2]

***Abstract*—Digital forensics field faces new challenges with emerging technologies. Virtualization is one of the significant challenges in the field of digital forensics. Virtual Machines (VM) have many advantages either it be an optimum utilization of hardware resources or cost saving for organizations. Traditional forensics' tools are not competent enough to analyze the virtual machines as they only support for physical machines, to overcome this challenge Virtual Machine Introspection technologies were developed to perform forensic investigation of virtual machines. Until now, we were dealing with persistent virtual machines; these are created once and used many times. We have extreme version of virtual machine and that is disposable virtual machine. However, the disposable virtual machine once created and are used one time, it vanish from the system without leaving behind any significant traces or artifacts for digital investigator. The purpose of this paper is to discuss various disposable virtualization technologies available and challenges posed by them on the digital forensics investigation process and provided some future directions to overcome these challenges.**

***Keywords*—Digital forensics; digital investigation; disposable virtual machines; light weight virtual machine; Microsoft sandbox; QEMU; qubes**

I. Introduction

Digital forensics is the process with four basic phases: collection, examination, analysis and reporting. During collection phase, data related to a specific event is identified, collected, and its integrity is maintained. Examination phase uses forensic tools and techniques as well as manual processes to identify and extract the relevant evidences from the collected data. Analysis phase deal with analyzing the results of the examination phase to generate useful information related to the case. Final phase generates reports of evidence from the results of the analysis [1]. A virtual machine (VM) is a tightly isolated software container with an operating system and applications inside. VM is self-contained and independent. Multiple VMs on a single physical machine with different operating systems and applications to run on just one physical server, or host. Hypervisor is the software layer, which decouples the virtual machines from the host and dynamically allocates and manages the computing resources to each virtual machine as per requirement [2]. Forensic investigation of virtual machines is challenging task if in a case virtual machine is subject of crime investigation, obtaining the image of the physical drive will not result in significant evidence since the virtual hard drive holds the evidence and more over vulnerabilities and attacks that affect the physical drive will have same effect on virtual environment. Analyzing multiple virtual machines using traditional tools of forensics is not possible. Virtual Machine introspection is the technique to monitor a virtual machine through hypervisor or a privileged VM, where the evidence collected without affecting the target VM [3]. Virtual machines created using oracle virtual box can be recovered using autopsy and other tools but VMs which were deleted using destroy command cannot be recovered [4]. The goal of this paper is to explore the disposable virtual machines and challenges posed to the digital forensics practitioners. Next section will discuss the virtualization technologies. Section 3 explores the disposable virtual machine technologies. Section 4 explores the challenges and roadblocks introduced by disposable virtualization to digital forensics. Section 5 will discuss current solutions to the issues related to disposable virtualization. We conclude with possible research directions to overcome these challenges.

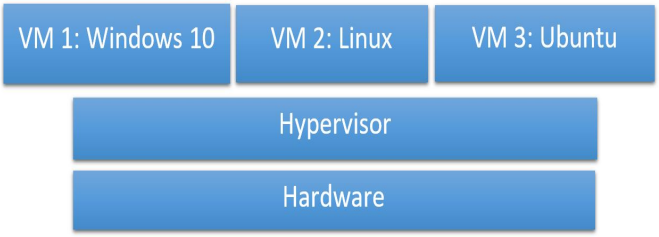


Fig. 1. Type-1 Hypervisor.

II. Virtual Machines

Virtualization technology enables utilization of resources in an effective way, reduces maintenance and security cost for the end-users. Virtual machine runs up on hypervisor. Hypervisors are of two types, one, which directly operates on physical hardware and does not require operating system, is called type-1 hypervisor, often called as "bare metal" hypervisors, examples include Citrix, Xen Server, ESXi from VMware, and Microsoft's Hyper-V. Layerd architecure of type-1 hypervisor illustrated in Fig. 1.

Second type of hypervisor rests upon operating system known as type-2 hypervisor. Most popular type-2 hypervisors are VMware, Virtual Box, and Parallel Desktop for MAC OS.

* Corresponding Author : Sultan Ahmad

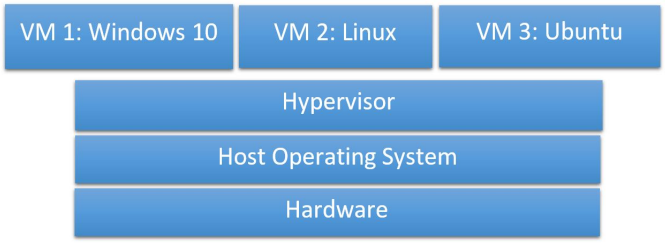


Fig. 2. Type-2 Hypervisor.

Type1 hypervisors provide greater performance and security and there is no overhead task for hypervisor to interact with host operating system. Type-2 hypervisor runs as an application on top of the host operating system (OS), it gives convenience to the individual users who intend to emulate a different operating system other than their OS, example: windows users can install Linux on virtual machine [5]. Fig. 2 shows the type-2 hypervisor's architecture. VMware files like vmdk file is virtual hard disk and vmem file is paging file act as primary memory RAM[6]. Oracle Virtual Box hypervisor also maintains such files, for each virtual machine there is a machine folder, inside machine folder vmname.vbox file and vmname.vdi , vdi format file for disk image, and Log files folder and a snapshot folder. These specific files of virtual machine collected from the host machine, to conducted investigation on virtual machine [7].

A. Virtual Machine Forensics

VM Forensics is similar to traditional digital forensics in many ways but at the same time, it introduces new pitfalls. Forensic approaches for virtual machines are many. Simplest form of forensics investigation of virtual machine starts with acquiring disk image of host computer on which virtual machines are running, after acquiring disk image files are extracted for the respective Virtual machine manger. Along with VM's files network logs and host operating system's registry also extracted. Disk image acquisition has to be done with utmost care, to preserve the integrity to ensure the legal admissibility of the evidence. There are standard procedures and guidelines for digital evidence acquisition approved by the Association of Chief Police Officers of the UK (ACPO), ISO Standard 27037, U. S. Department of Justice Office, and the EU publication Guidelines on Digital Forensic. First, the machine is powered off by disconnecting power supply. Then the hard disk drives or solid-state drives disassembled from the suspect machine. Extracted disk drive is write protected with write blocker kit. Disk drive then connected to forensic machine to create a duplicate image of disk drive using specialized tools such as dd, FTK imager and "encase", etc. Disk image acquired from previous step is used for analysis. In case of VM disk image there are two approaches, first is resuming the suspended virtual machine on corresponding virtual machine manager. Second approach is to create the snapshots of virtual machine.In case of resuming the suspended virtual machine VM disk files vmdk, or vdk or vhd files and other files related to virtual machine are restored, down side of this approach is during resuming process VM files may change and integrity of the evidence is compromised. While snapshot of VM used for forensic analysis, there will be no changes on state of the HDD. Forensic analysis tools; Encase, FTK supports the conversion of virtual disk image files (.vmdk, .vdi) to raw dd format files [8]. Virtual machine introspection technique uses virtual machine manager to view inside virtual machine, to track and view virtual machine state. VMI can inspect and view VM-memory, processor, installed Operating systems, applications and services. Evidence Search through injected code. This strategy is inspired by code injection attacks. Which uses vulnerabilities to inject malicious code in to applications and kernel to control and corrupt the system [9].

TABLE I. Disposable Virtual Machines.

Disposable VM	Hypervisor	Type
Microsoft Sadbox	Microsoft Hypervisor	Type 2
Qubes Disposable VM	KVM, Xen	Type 1
Virtual box Nested VM	Virtual Box	Type 2
Shade SandBox	Microsoft Hypervisor	Type 2
QEMU	Xen, KVM, Hax	Type 2
Bitbox	Virtual Box	Type 2

III. Disposable Virtual Machines

Disposable virtual machine is the lightweight virtual machine, created instantly and it will be disposed when it is closed. Disposable VMs commonly used to host single application, such as web browser, viewer, editor and suspicious applications. This concept of single use virtual machines also adopted by various operating systems. In Table I, the few popular disposable Virtual machine managers are listed.

A. Microsoft Windows Sandbox (WSB)

Microsoft Windows sandbox runs applications in isolation. Secure execution of application in sandbox environment does not affect the host operating system. New instance of sandbox created each time and disposed as soon as it is closed. Preinstalled applications in host operating system are not accessible in sandbox environment instead explicit installation of application is required. Sandbox uses hardware virtualization for kernel isolation. Windows Sandbox is a new lightweight disposable desktop environment. Which runs application in isolation. Windows 10 pro and enterprise editions include sandbox environment. As soon as sandbox is closed, applications and residual files, and data related to that particular sandbox deleted permanently. Every time you start a Windows Sandbox, it is as clean as a brand-new installation of Windows. Windows 10 operating systems has all required files pre-loaded to run the sandbox. It is disposable nothing persists on the host device as soon as you close the sandbox.Windows Sandbox (WSB) gets the dynamically generated base image with its own directory structure as host operating system, except the mutable files are copied in to WBS directory structure. Immutable files of host operating system can be accessed through links. Efficiency of the windows sandbox achieved by following: process scheduling integrated with kernel scheduler. Smart memory management where memory pages are allocated to WSB and Host operating system on demand, there is no fixed chunk of memory for WSB, it gives more flexibility and improves efficiency overall. virtual GPU enables dynamic utilization of graphics processing. Windows sandbox architecture

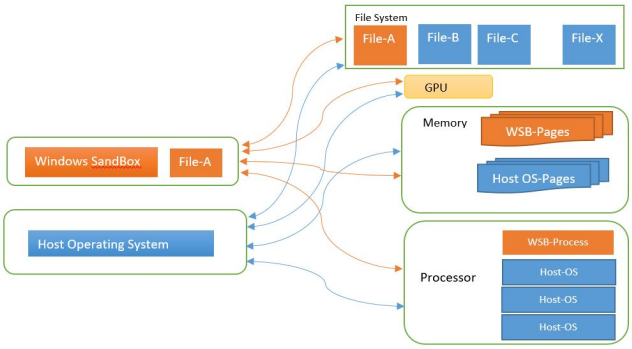


Fig. 3. Windows SandBox Architecture.

is illustrated in Fig. 3. To use windows sandbox you must start the sandbox first and copy the executable file you wish to run from the host file system and paste the executable file in sandbox. Once the file copied, you can run it as a normal application. Windows Sandbox gives two options; one is to run a full desktop in sandbox. Second option is just the application in sandbox and as known as rails. Sandbox has many advantages over tradition virtual machines creates a virtual machine and installs complete operating systems where resources are shared among host operating system and virtual machines. In case of windows, only few files used for sandbox from host file system it is dynamically generated image. Memory management is dynamic based on payload system allocates memory to the sandbox. Process scheduling is integrated where sandbox and host systems are managed together . Windows sandbox is secure as it runs on a separate kernel that provided by Microsoft's hypervisor keeping it isolated from the host kernel. Virtualization in case of sandbox is hardware-based illustrated in Fig. 1. Thus to implement type-1 hypervisor host system must support virtualization, which can be enabled or disabled from BIOS of the host system. Any malicious code will not affect the host kernel and will not persist as soon as sandbox is closed. WSB can be accessed remotely from server where Sandbox is created in two modes 1.WSB with full desktop 2.WSB Rails in Rails a specific application is launched on sandbox it is similar as Application VM. Remote clients can access and launch the WSB from server, once it is closed no files or changes are saved in host server[10].

B. Qubes Disposable VM

Qubes OS developed with focus on Security through isolation approach. Virtualization is based on Xen hypervisor. Domains created with different security levels, which runs on virtual machine. Work domain is more secure than Shopping domain. Dom0 is the administrative domain it can access all the hardware directly, such as graphics devices, input output devices like keyboard and mouse. This administrative domain manages the virtual disks of the other VMs, it stores these virtual disk images on its file system. Disk space saved by storing virtual disk on same file systems and accessed in read only mode. Qubes allows users to launch disposable VM directly from dom0's start menu or from an AppVM you have to choose open with disposable VM. In disposable VM you can work with untrusted files without compromising other Virtual machines. Disposable VMs created using Disposable VM Template. Disposable VMs created with these templates has its own user file system, one for each disposable VM. Qubes R4.0 has multiple Templates and default template for disposable VM is fedro-xx—dvm(xx here refers to version number[11].

C. VirtualBox

Oracle VirtualBox is an open source type 2 hypervisor for virtualization of window and Linux operating systems from Oracle Corporation. Creation and management of guest virtual machines is very much user friendly. Intel VT-x and AMD-V hardware-assisted virtualization is supported on VirtualBox. It supports nested virtualization that is one of the challenges for digital forensics experts [7]. Nested virtual machines runs on hypervisor which is on top of other virtual machine, this stacking of hypervisor recursively increases overhead but at the same time provides extra layer of security and decouples the VM from physical host [12]. Eventually it comes with extra overhead for digital forensic investigation.

D. Shade Sandboxie

Shade Sandboxie is an application based sandboxing. It creates isolated environment to execute suspicious code. Such an environment is used to track and notice code behavior and output activity, it creates functional layer of network security against ATPs and other cyber threats. Applications run inside simulated virtual environment without hardware virtualization support. Running malicious code and browsing websites with potential threats will not affect the host Operating System [13].

E. QEMU (Quick Emulator)

QEMU is the hosted virtual machine monitor it operates in different modes. System emulation mode where it emulates hardware including processor, peripheral devices. In user mode, it runs programs using different instruction set rather than its instruction set by cross-compilation and cross debugging. KVM hosting mode, QEMU emulates hardware but guest operating system runs on KVM. XEN hosting mode, here also QEMU emulates hardware and XEN run the guest operating systems [14].

F. BitBox

BitBox is secure firefox encased in virtual machine with linux OS on oracle virual box. Only drawback of this is the setup, which takes 2GB of disk space. Developed by German cyber Security Company Rohde and Schwarz to prevent cyber-attacks such as APTs, Zero-day exploits and Ransomwares[15].

IV. Challenges Posed by Disposable Virtual Machines in Digital Forensics

That, in essence, attackers can start a disposable VM to carry out their act and close the disposable VM, which leaves no traces for forensics expert. Existing Virtual machine forensic techniques are not going to yield significant results. The disposable virtual machines not designed with digital forensics

and evidence integrity in mind, instead the objective was to completely isolate applications from host operating system and leave a pristine system without leaving any traces behind. However, not any significant work has been done in disposable virtual machine forensic. We could not find any substantial information about disposable virtual machine or lightweight VM forensics. Forensic investigation begins with identifying the system, which contains potential evidence or involved in suspicious activity. First step is to identify the incident and next is to acquire evidence to prove the incident. When it comes to disposable virtual machines, no traces are left. The very nature of disposable virtual machines architecture is the main challenge in data identification and subsequent collection of evidence. Mostly no artifacts left after closing disposable virtual machines. Possible solution could be capturing the sandbox or the disposable virtual machine instances while they are active other possible solution is to perform data carving from memory dumps log files of hypervisor. In presence of hypervisor, it is difficult to take, the memory dump of the physical memory it is difficult to extract the data from memory reserved fur virtual machine monitors. One possible way to use memory acquisition tools like volatility, Rekall and Layout Expert [16].It might be able to analyze virtual machine processes running on the machine even after capturing memory dumps it is difficult to analyze the memory dump for virtual machine data. Here we use the standard forensic investigation steps to discuss the challenges posed by disposable VMs at each stage. Stage of forensic investigation are as follows: 1. Forensic Image creation 2. Identification and Recovery 3. Analysis 4. Presentation and Documentation[17].

A. Forensic Image Creation

Disk image of suspected system is created from physical machine. At this stage, integrity of the image created must be preserved. This is performed using tools like DD, DDRescue, Encase and Photorec etc.[18][19]. Investigator never uses the original disk to conduct investigation; instead, image of the disk used to conduct analysis and further investigation. This image used to collect the information about virtual machine and hypervisor used. Information included execution time logs, temporary files, snapshots and Internet activity log files etc. Therefore, investigator must collect the image carefully without tampering its integrity to extract vital information. Write blockers are used to prevent accidental writes on to the original disk. MD5 hashing is one of the method to ensure the integrity. Forensic tools allow us to complete this task by mounting disk image for further analysis of the Virtual machines and Hypervisor. Graphical user interface such as Dymanage and AIR are developed for DD find DD rescue. In case of disposable virtual machines, data is not persistent so it is not possible to create disk image of disposable virtual machines.

B. Identifcation and Recovery

At first, host machine is analyzed to find the traces of virtual machine in hypervisor. Host operating system maintains log files, which lead to extract traces of virtual machine. Windows operating system maintains registry entries, prefetched files, shared DLL, log files, thumbnails, icons, temporary files, and system event logs etc. that can prove the virtual machine existence in host computer. Even after files are deleted most of the time operating system do not completely delete the files instead removes the file reference from master file table. Specifically for large size, files such as virtual machine files still exist in the disk. Each of these files can be extracted from the unallocated space of the secondary disk. Data recovery tools like best disc, handy recovery and R-studio etc. commonly used to recover data from the disk image.

TABLE II. DISPOSABLE VM CHALLENGES.

Investigation Stage	Challenge
Image creation	No persistent files of disposable VM exist on disk drive
Information identification	Host OS or the Hypervisor do not maintain activity logs of disposable VM
Analysis	Snapshots or .vdi files are not available
Presentation	No specific format of reporting is available

C. Analysis

Virtual machine analysis: regular virtual machines can be Analyzed by mounting it as disk drive or by accessing it through a hypervisor. In case of disposable virtual machines, it is not possible; files related to disposable VM are deleted. Virtual machine files could be recovered by identifying its format based on hypervisor. Files with extensions like .VDI, .VMDK is used by popular hypervisors [20]. Other options to investigate the virtual machine is by picking a snapshot of VM and further analyzing snapshot to extract the vital information that can be presented as evidence. This provision of snapshot is not available for disposable virtual machines. Only option left for disposable VM is to analyze host operating system log files, registry entries, etc.

D. Presentation and Documentation

Documenting and presenting the evidence found during investigation is the final stage of forensic investigation. Evidence includes time stamps, who accessed and when accessed data or performed an activity. Forensics tools have their proprietary format of reports. There are no specific forensic tools for disposable VM forensics, so there exists no specific reporting formats for disposable VM. It is preferable to use the same reports as virtual machine. In Table II we have presented the challenges posed by disposable VM at every stage of forensic investigation.

V. CONCLUSION

In this paper, the investigators have explored challenges posed by lightweight VM to the digital forensics experts at every stage of digital forensics investigation. We discovered that there is not much research done in disposable VM forensics. These challenges needs to be addressed by conducting experiments on disposable VM. One of the possible thing is to compare the complete system image before and after running disposable virtual machine on various platforms and in this way we find possible traces or changes in the system.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia.

REFERENCES

[1] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to integrating forensic techniques into incident response," *NIST Special Publication*, vol. 10, no. 14, pp. 800–86, 2006.

[2] VMWare.com, *VMware*, 2020 (accessed October 20, 2020). [Online]. Available: https://www.vmware.com/solutions/virtualization.html

[3] J. Poore, J. C. Flores, and T. Atkison, "Evolution of digital forensics in virtualization by using virtual machine introspection," in *Proceedings of the 51st ACM Southeast Conference*, 2013, pp. 1–6.

[4] E. Wahyudi, I. Riadi, and Y. Prayudi, "Virtual machine forensic analysis and recovery method for recovery and analysis digital evidence," *International Journal of Computer Science and Information Security*, vol. 16, 2018.

[5] P. Tobin and T. Kechadi, "Virtual machine forensics by means of introspection and kernel code injection," in *Proceedings of the 9th International Conference on Cyber Warfare & Security: ICCWS*, 2014, p. 294.

[6] S. Lim, B. Yoo, J. Park, K. Byun, and S. Lee, "A research on the investigation method of digital forensics for a vmware workstation's virtual machine," *Mathematical and computer modelling*, vol. 55, no. 1-2, pp. 151–160, 2012.

[7] Virtualbox.org, *VirtualBox*, 2020 (accessed October 20, 2020). [Online]. Available: https://www.virtualbox.org/manual/ch10.html

[8] M. Hirwani, Y. Pan, B. Stackpole, and D. Johnson, "Forensic acquisition and analysis of vmware virtual hard disks," 2012.

[9] P. Tobin, N.-A. Le-Khac, and T. Kechadi, "Forensic analysis of virtual hard drives," *Journal of Digital Forensics, Security and Law*, vol. 12, no. 1, p. 10, 2017.

[10] Microsoft, *Windows Sandbox*, 2020 (accessed October 15, 2020). [Online]. Available: https://docs.microsoft.com/en-us/windows/security/threat-protection/windows-sandbox/windows-sandbox-overview

[11] Q. OS, *DisposableVMs*, 2020 (accessed October 15, 2020). [Online]. Available: https://www.qubes-os.org/doc/disposablevm/

[12] B. Kauer, P. Verissimo, and A. Bessani, "Recursive virtual machines for advanced security mechanisms," in *2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 2011, pp. 117–122.

[13] shadesandbox.com, *Shade Sandbox*, 2020 (accessed November 22, 2020). [Online]. Available: https://shadesandbox.com/blog

[14] qemu.org, *Quick Emulator*, 2020 (accessed November 22, 2020). [Online]. Available: https://www.qemu.org/documentation/

[15] https://www.rohde schwarz.com, *Browser In The Box*, 2020 (accessed November 22, 2020). [Online]. Available: https://www.rohde-schwarz.com

[16] T. Wu, F. Breitinger, and S. O'Shaughnessy, "Digital forensic tools: Recent advances and enhancing the status quo," *Forensic Science International: Digital Investigation*, vol. 34, p. 300999, 2020.

[17] S. R. Selamat, R. Yusof, and S. Sahib, "Mapping process of digital forensic investigation framework," *International Journal of Computer Science and Network Security*, vol. 8, no. 10, pp. 163–169, 2008.

[18] N. Reddy, "Linux forensics," in *Practical Cyber Forensics*. Springer, 2019, pp. 69–100.

[19] S. Widup, *Computer forensics and digital investigation with EnCase Forensic v7*. McGraw-Hill Education Group, 2014.

[20] H. Riaz and M. A. Tahir, "Analysis of vmware virtual machine in forensics and anti-forensics paradigm," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*. IEEE, 2018, pp. 1–6.

Priority-Mobility Aware Clustering Routing Algorithm for Lifetime Improvement of Dynamic Wireless Sensor Network

Rajiv R. Bhandari[1], Dr. K. Raja Sekhar[2]
Research Scholar[1],
Department of Computer Science and Engineering[1,2],
K L Education Foundation,
Vaddeswaram -522502, Guntur Dist., A.P, India

***Abstract*—Wireless sensor network with mobility is rapidly evolving and increasing in the recent decade. The cluster and hierarchical routing strategy demonstrates major changes in the lifespan of the network and the scalability. The latency, average energy consumption, packet distribution ratio is highly impacted due to a lack of coordination between cluster head and extreme mobile network nodes. Overall efficiency of highly mobile wireless sensor network is reduced by current techniques such as mobility-conscious media access control, sleep/wakeup scheduling and transmission of real-time services in wireless sensor network. This paper proposes a novel Priority-Mobility Aware Clustering Routing algorithm (p-MACRON) for high delivery of packets by assigning fair weightage to each and every packet of node. To automatically decide the scheduling policy, reinforcement learning approach is integrated. The mixed approach of priority and self-learning results into better utilization of energy. The experimental result shows comparisons of slotted sense multiple access protocol, AODV, MEMAC and P-MACRON, in which proposed algorithm delivered better results in terms of interval, packet size and simulation time.**

Keywords*—Cluster; routing; sleep scheduling; priority; reinforcement

I. Introduction

Recently wireless sensor networks are rapidly growing and shifting their paradigm from static to dynamic. Such dynamic changing environment are highly impacted by traditional MAC protocol and sleep scheduling algorithm. Newly developed mobile sensors are advance and much more constrained related to mobility [1] [2], routing protocol, scheduling and clustering [3], [4], [5]. One of the most important features to consider in fast-growing WSN is node mobility and data transmission in real time. A Preemptive Priority-Based Data Fragmentation Scheme proposed [6] where high priority packets and low priority packets are handled. FROG-MAC focuses on fragmentation of packets so that higher priority packets need not to wait for longer time while low priority packets are transferred. The limitation of this scheme is interference of high priority packets while low priority packets are in transmission. Fasee Ullah et. all [7] proposed TRIP-ECC protocol which mainly classify priority of data into four different categories like usual data, on demand data, emergency data of low threshold reading and high threshold reading. This classification works well with Wireless body network but did not work well in heterogeneous network where real time delivery and fair weightage for all nodes. To handle the mobility and scalability of network Mahdi Zareei et al. [5] [8] has done an extensive survey of mobility aware MAC protocol. Based on their survey the paper categorized MAC protocol Scheme in four major categories: General active/sleep time, Slotted TDMA based MAC protocol, preamble sampling MAC protocol and hybrid MAC protocol. Mobility aware protocol is well described with their pros and cons in this paper. MEMAC algorithm is proposed by Bashir Yahya et al. [9] is a hybrid MAC protocol to handle the mobility of nodes with cluster creation, shift, leave, join operation very well. The main constrain of this protocol is handling mobility in large and scalable network. Dayong ye and Minjie Zhang implemented self-adaptive sleep/wakeup scheduling algorithm. This algorithm avoids use of duty cycling to overcome the problem of packet deliver and energy saving. This algorithm uses the concept of Reinforcement learning to achieve better results and algorithm indeed gives best results. The authors didn't test the algorithm on any of cluster routing protocol. This work we have extended in our P-MACRON.The primary role of proposed algorithm in further section is to handle the traffic in mobile network by allocating fair and equal weightage all node. During the data transmission, the node priority is validated by extracting the details from the Packet Header. The remaining paper is organized as follows the background knowledge is described in Section 2, Section 3 specifies proposed work, implementation of work is mentioned in Section 4, results are discussed in Section 5 followed by conclusion in Section 6.

II. Background Knowledge

The proposed work is extension of mobility aware clustering routing MACRON algorithm proposed in [10] [11]. This algorithm wisely chooses the cluster head and its members based on probability distribution functions and iterative calculation. Unlike conventional algorithms, MACRON manages node mobility effectively by performing leave join operations iteratively. The MACRON algorithm operates in three phases: 1) Network creation MACRON Algorithm, 2) Self-healing Scheduling MACRON Algorithm, 3) Self-healing scheduling approach using Reinforcement Learning. In this section, these three phases are described in detailed.

A. Network Creation

The system consists of base station and sensor nodes. The base station initiated clustering process based on node deployment and coverage area. Clustering is executed by an iterative process with probability distribution function. The initial probability of each node is determined based on the one-hop connectivity distance to reach all the covered sensor nodes. By applying LEACH method with energy parameter the updated probability for each node is calculated. Normalize the estimated probability by computing the ratio of the current probability value to the cumulative probability value of all sensors [1] [10]. The node with max probability is selected as Cluster head. The node with one hop distance are selected as members for cluster head. The same process will continue until all nodes in the network are reached to the end [12], [13] .

Algorithm 1: Network and Cluster Head Creation

1 Create location table by observing location x,y and energy of node.;
2 **for** $i = 1, \ldots, n$ **do**
3 $\quad nodex(i) \leftarrow x_i$;
4 $\quad nodey(i) \leftarrow y_i$;
5 $\quad nodee(i) \leftarrow energy$;
6 **end**
7 To form the cluster table unit area, average cluster node and number of clusters are calculated

$$C_{area} = \frac{N}{Max\_x * Max\_y}$$

$$Avg_{Clust\_node} = C_{area} * P_i * node^2$$

$$Clust_{num} = \frac{number\ of\ nodes}{Avg_{Clust\_node}}$$

8 using one hop distance the cluster are formed ;
9 The node having max probability will be selected as cluster head CH_j;
10 Cluster are refined using probability distribution and LEACH technique and cluster list is created list $CH_j \leftarrow nd(i)$;

B. Algorithm: Join, Shift and Leave Operation

Cluster member chooses tentative cluster head based on shortest reachability and sends leave message to old cluster head and join message to newly elected cluster head. If one cluster head receives CH announcement from other overlapping cluster head, then need to find cluster probability. The node with the greater probability retains the head position so the lower probability CH performs the cluster shifting process by sending the shifting message to both CM and overlapping CH, then cluster head assigns slot to members [14] [15]

C. Self Healing Scheduling Approach using Reinforcement Learning

Self-healing scheduling approach using Q-learning algorithm decides when to transmit the packets in sub time-slot. This algorithm works with rewards and penalty so that

Algorithm 2: Join, Shift and Leave Operation

1 Cluster member CM choose cluster head $Curr\_CH$ based on shortest reachability;
2 **if** $CM \notin Curr\_CH$ **then**
3 $\quad$ send leave message to Old_CH and join message to $Curr\_CH$.
4 **end**
5 **if** $CMnot \notin Curr\_CH$ *and receives message from* $Curr\_CH$ **then**
6 $\quad$ perform cluster shifting by sending leave and join message.
7 **end**
8 CH received announcement from overlapping $Curr\_CH$ then $Curr\_CH$ checks cluster probability.

all nodes will select transmission of packets with proper prediction in future. This algorithm works with sleep and active states by adjusting the idle state of node as it consumes lot of energy[3][16]. Initially the learning rate ζ, and δ are determined by observing the how much amount of prior information will be override by the recent set of actions. The set of actions considered here is $a \in \{transmit, listen, sleep\}$

Generally it takes the value between $[0, 1]$, Here, 0 means that the prior information will be retained as it is and 1 indicates that it will positively identify the prior information. The discount factor γ also takes the value between $[0, 1]$, the value nearer to 0 indicates the node is naive and it deals with the recent actions only and, the discount factor value approaching to 1 indicates that the node is promising with high reward value.

For reinforcement learning approach, the policy defined in [17] [18] is adopted. By considering this policies the payoff, average payoff will be decided and the transmission slot will be normalized.

Algorithm 3: Self-healing scheduling approach using Reinforcement Learning

1 Initialize the learning rate ζ, and δ, discount factor γ by observing the current action $a\_current$;
2 According to rule, select set of available actions ;
3 Decide Payoff, Next State and time slot of when to transmit packet;
4 calculate the average payoff;
5 Decide the probability of selection of subslot of transmitting packet;
6 Normalize the slot;

III. Proposed Work

The proposed work is demonstrated in three sections. The detail architecture is shown in Section 1, Section 2 focuses on how reinforcement learning is integrated in proposed work.The P-MACRON algorithm is described in Section 3.

A. Architecture of Proposed Algorithm

The Architecture of sensor network shown in Fig. 1 is created by consideration of Node 0 as Base node and base

station initiate the clustering process on nodes deployment and coverage area of sensors. The network will perform operations such as joining another network, switching from one network to another and finding nodes in the network about to die. The node with the greater probability retains the head position so the lower probability CH performs the cluster shifting process by sending the shifting message to both CM and overlapping CH [19].

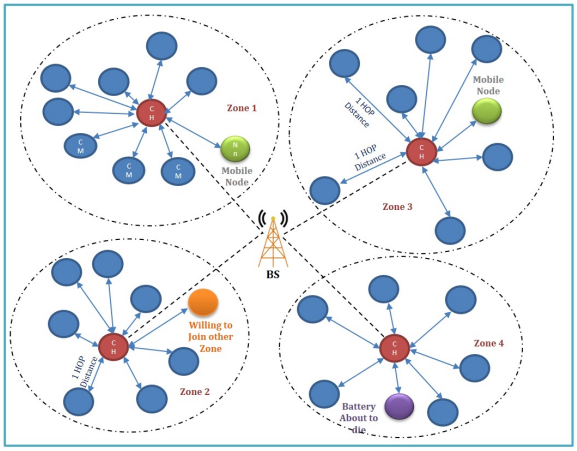


Fig. 1: System Architecture with Join, Leave, Shift Operation.

B. Reinforcement Learning Integration with Proposed Algorithm

Reinforcement learning enables the nodes to learn best possible actions in order to take best decision with previous experiences as shown in figure 2.

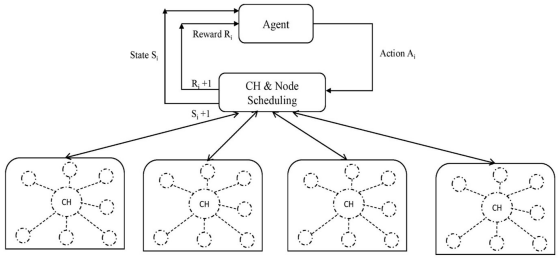


Fig. 2: Self-healing Sleep/wakeup Structure using Reinforcement Learning Algorithm.

C. P-MACRON Algorithm

The proposed work contributes in maximization of network lifetime for dynamic wireless sensor network and giving fair chance to all nodes in the network to transmit the data. The three objectives of algorithm are:

- MACRON algorithm for minimum energy consumption that works effectively and proficiently in dynamic mobile wireless area network.
- Self-healing scheduling using reinforcement learning approach for long life of wireless sensor network.
- A new priority based hybrid approach for improving lifetime of Wireless Sensor Network to deliver real-time data by assigning fair weightage to every node.

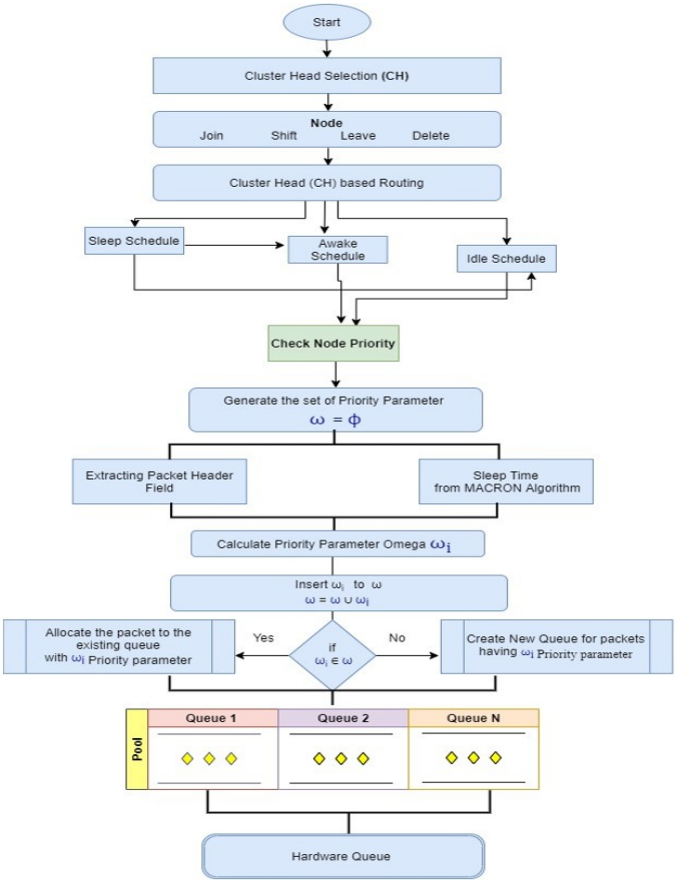


Fig. 3: Flow of Priority-MACRON Algorithm.

The third objective of the algorithm is proposed here to elongate the lifetime of wireless sensor network. A main research theme in dynamic wireless sensor networks is the nature of real-time data transmission and equal chance for every node. P-MACRON maintains priority queue $[PQ1, PQ2, PQ3 \ldots PQn]$ and tuning factor ω to assign fair weightage to all nodes. Proposed P-MACRON extract packet header parameters and consider scheduling value like sleep and active. Self-healing scheduling algorithm adjust their idle state with sleep and active state this will help in fast transmission of packets. Each node maintaining its $sleep$ and $active$ state and communicate its schedule with cluster head locally. All cluster heads will dynamically maintain their cluster table with schedules communicated by nodes.During the data transmission, the node priority is validated by extracting the details from the Packet Header. The packet header contains the information about the parameters such as the average number of nodes distributed in the clusters, number of packets in the buffer, sleep time for each sensor, sensors connectivity with its neighbors and the mobility of the sensors [20] [21]. The weighted fair queuing model is applied to decide the priority based on the uniformly distributed weight values. The cumulative value of

the weight is maintained to the unit value and the weighted value is computed for all parameters related to the priority value. The weighted priority value of the data packet is used to decide the packet queue which is used to buffer the packet. If the priority is not available in the queue then it is included in the queue pool to complete the data transmission. If the priority is available in the queue then the same pool is used to complete the data transmission. Based on the selected queue pool the transmission priority is assigned to each packet. In the Mac layer, the packet transmission is performed based on the assigned priority. Fig. 3 is explaining the overall flow of P-MACRON Algorithm.

Algorithm: P-MACRON
Input: list(CH_j) with nodes $nd(i)$
Output: Priority Queue Portions $[PQ_1, PQ_2 \dots PQ_n]$, $\{nd(1), nd(2) \dots nd(i)\} \in PQ_i$
Method:
Generate the set of Priority Parameter ω;
repeat
  for each packet of $nd(i)$;
  extract packet header (PH_i);
  observe the transmit time $T_st(i)$;
  check the status $[nd(i)]$;
  if *status* $[nd(i) == "Sleep"]$ **then**
    $\Delta(i) = Sleep\_Time(i)$
  else
    $\Delta(i) = 0$
    $\omega(i) = PH(i) + T_{st}(i) + \triangle(i)$
    $\omega = \omega \cup \omega(i)$
  end
until *List* $CH(i) \neq \emptyset$;
$k = 1$
foreach $nd(i) \in list(CH_j)$ **do**
  $Extract\ \omega_i\ for\ nd(i)$
end
if PQ_k *is available with Index* ω_i **then**
  $insert\ nd(i) \rightarrow PQ_k$
else
  $k = k + 1$;
  $create\ PQ_k$ with Index ω_i;
  $insert\ nd(i) \rightarrow PQ_k$;
  foreach $i := 1$ *to* k **do**
    arrange the Priority Queue Partition $PQ_i \leq PQ_{i_{+_1}}$
  end
end
return a

IV. Environment Setup and Results

This section proposed more stable and consistent result over slotted sense multiple access algorithm. The framework uses $NS2.34$ to conduct performance study to analyses P-MACRON and SSMA efficiency [22], and to evaluate P-MACRON's feasibility. The network size of 500 by 500 square meter is considered and 101 nodes are randomly deployed where node 0 is working as Base station. The simulation results of P-MACRON in comparison with SSMA are based on three main parameters like interval, packet size and simulation time. The simulation parameters are mentioned in table I.

TABLE I: Simulation Parameters

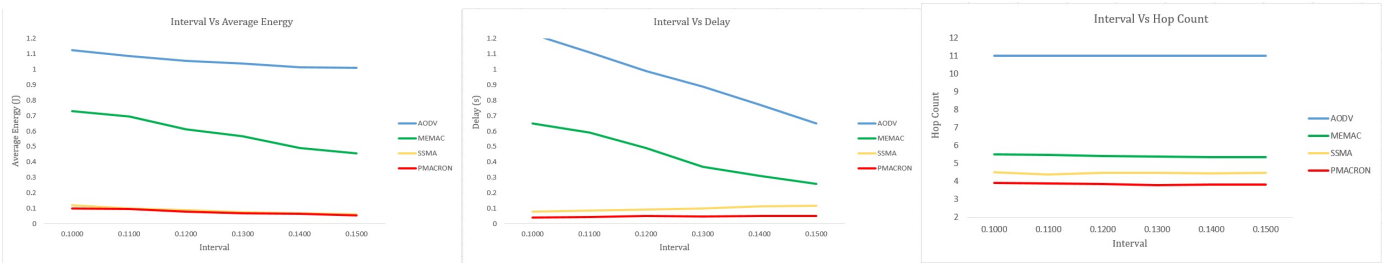
Particular	AODV	MEMAC	SSMA	P-MACRON
Node	101	101	101	101
Network Size	500 * 500 sq mtr			
MAC	AODV	MEMAC	SSMA	802.11
Radio Range	$3.652e-8$			
Simulation Time	200s			
Traffic Source	Sense Traffic			
Packet Size	50 bytes			
Mobility Model	Random Mobility			

Fig. 4 and Fig. 5 are showing comparison of P-MACRON with SSMA, MEMAC, AODV. Some of the observation for the performance of the algorithm are as follows:

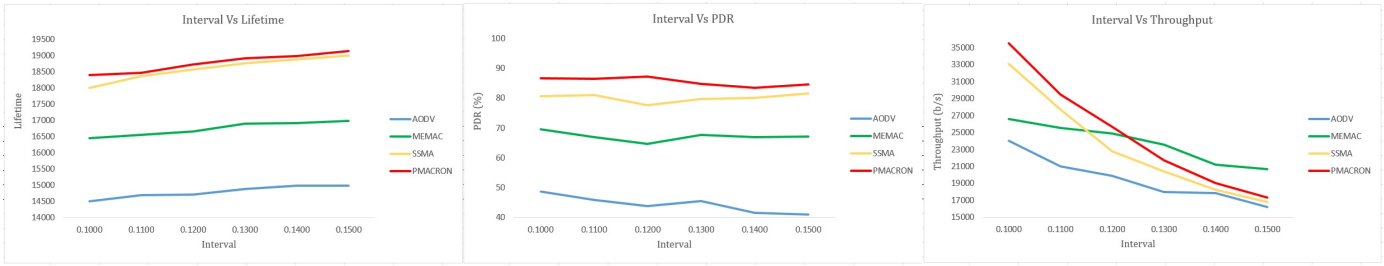
- The most significant efficiency metric for wireless sensor networks is energy consumption. The average energy consumption for AODV, MEMAC, SSMA under variable interval and packet size for mobile network is shown figure. P-MACRON outperforms AODV, MEMAC and SSMA as mobility increases.
- The percentage of packet delivery of P-MACRON to base station is significantly promising over AODV, MEMAC and SSMA.
- The interval and packet size changes from 0.1000 to 0.2000 and 20 to 40 respectively. Under the mobile network scenario, the AODV, MEMAC have high delay but still SSMA and P-MACRON shows better performance. If we compare SSMA, MEMAC and AODV [23] as PMACRON adopts a clustering
- approach with one hop distance, the hop count needed to reach the destination is significantly fine.
- The Overall lifetime of network is main focus of proposed algorithm. Figure shows significant growth in improving lifetime of network over SSMA, MEMAC and base case AODV.
- The number of packets per second received at receiver is defined as throughput. The number of packets delivered at receiver end is quite good in P-MACRON and MEMAC over SSMA and AODV.

V. Conclusion and Future Work

Using the clustering and scheduling algorithm, assigning priorities has been a primary concern of the industrial sector and the defense sector. P-MACRON algorithm mainly focuses on assigning fair weightage to all packets of nodes for reliable and on time delivery of packets. As observed, P-MACRON lays the groundwork to group nodes at one hop distance into clusters with priority leading to energy-efficient routing and scheduling using machine learning algorithm has proven to be one of the most effective approaches for exclusive dynamic wireless sensor networks. P-MACRON algorithm works efficiently for extensively dynamic algorithm with distinguish method of cluster creation and self-healing scheduling algorithm using Q Learning Algorithm. Results from simulation

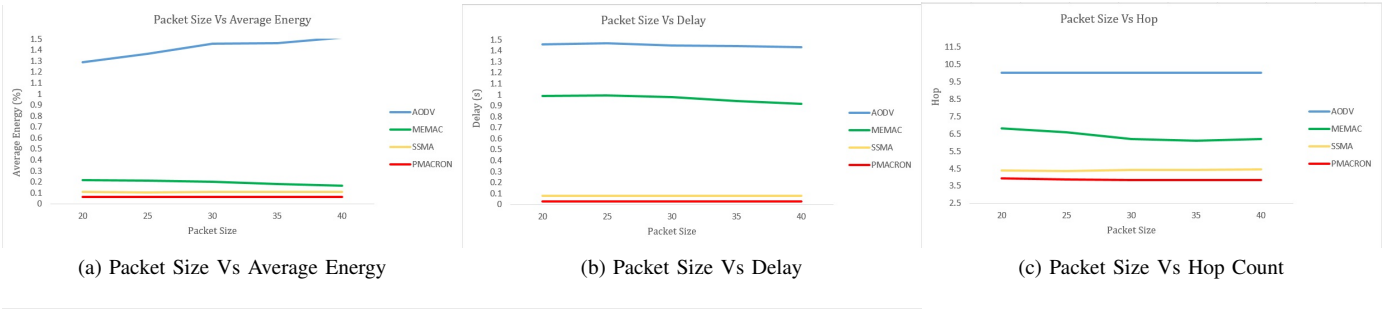


(a) Interval Vs Average Energy (b) Interval Vs Delay (c) Interval Vs Hop Count

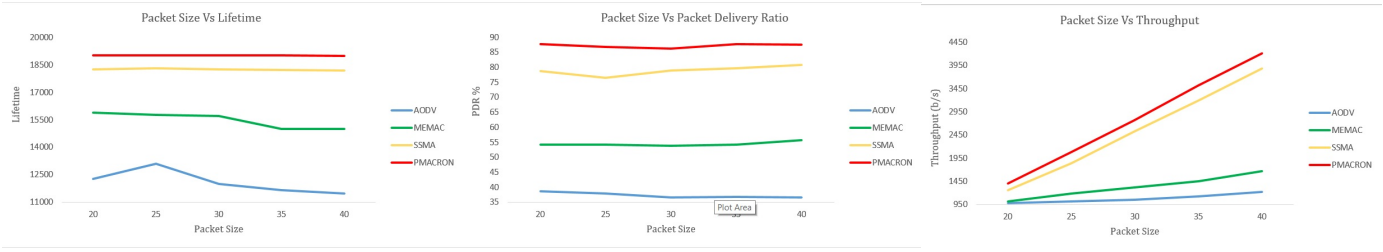


(d) Interval Vs Lifetime (e) Interval Vs PDR (f) Interval Vs Throughput

Fig. 4: Comparative Analysis based on Interval.



(a) Packet Size Vs Average Energy (b) Packet Size Vs Delay (c) Packet Size Vs Hop Count



(d) Packet Size Vs Lifetime (e) Packet Size Vs PDR (f) Packet Size Vs Throughput

Fig. 5: Comparative Analysis based on Packet Size.

reveal that P-MACRON performs much better in terms of Average Energy, delay, hop count, packet delivery ration and throughput than the SSMA, MEMAC and AODV algorithms to achieve high lifetime with high mobility. In future, we are planning to implement P-MACRON with topology independent network with more parameter such as multihop transmission, collision.

REFERENCES

[1] B. Yahya and J. Ben-Othman, "An adaptive mobility aware and energy efficient mac protocol for wireless sensor networks," in *2009 IEEE Symposium on Computers and Communications*. IEEE, 2009, pp. 15–21.

[2] M. Anusha and S. Vemuru, "Cognitive radio networks: State of research domain in next-generation wireless networks—an analytical analysis," in *Information and Communication Technology for Sustainable Development*. Springer, 2018, pp. 291–301.

[3] D. Ye and M. Zhang, "A self-adaptive sleep/wake-up scheduling approach for wireless sensor networks," *IEEE transactions on cybernetics*, vol. 48, no. 3, pp. 979–992, 2017.

[4] E. Srie Vidhya Janani and P. Ganesh Kumar, "Energy efficient cluster based scheduling scheme for wireless sensor networks," *The Scientific World Journal*, vol. 2015, 2015.

[5] M. Zareei, A. M. Islam, C. Vargas-Rosales, N. Mansoor, S. Goudarzi, and M. H. Rehmani, "Mobility-aware medium access control protocols for wireless sensor networks: A survey," *Journal of Network and Computer Applications*, vol. 104, pp. 21–37, 2018.

[6] A. A. Khan, S. Ghani, and S. Siddiqui, "A preemptive priority-based data fragmentation scheme for heterogeneous traffic in wireless sensor networks," *Sensors*, vol. 18, no. 12, p. 4473, 2018.

[7] F. Ullah, Z. Ullah, S. Ahmad, I. U. Islam, S. U. Rehman, and J. Iqbal, "Traffic priority based delay-aware and energy efficient path allocation routing protocol for wireless body area network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 10, pp. 3775–3794, 2019.

[8] P. Bansal, P. Kundu, and P. Kaur, "Comparison of leach and pegasis hierarchical routing protocols in wireless sensor networks," *International Journal on Recent Trends in Engineering & Technology*, vol. 11, no. 1, p. 139, 2014.

[9] L. Karim, N. Nasser, T. Taleb, and A. Alqallaf, "An efficient priority packet scheduling algorithm for wireless sensor network," in *2012 IEEE international conference on communications (ICC)*. IEEE, 2012, pp. 334–338.

[10] R. R. Bhandari and K. Rajasekhar, "Study on improving the network life time maximazation for wireless sensor network using cross layer approach," *International Journal of Electrical and Computer Engineering*, vol. 6, no. 6, p. 3080, 2016.

[11] M. El Ouadi and A. Hasbi, "Comparison of leach and pegasis hierarchical routing protocols in wsn," 2020.

[12] L. Farhan and R. Kharel, "Internet of things scalability: communications and data management," in *Modern Sensing Technologies*. Springer, 2019, pp. 311–329.

[13] B. Jan, H. Farman, H. Javed, B. Montrucchio, M. Khan, and S. Ali, "Energy efficient hierarchical clustering approaches in wireless sensor networks: A survey," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[14] A. Lohachab, A. Lohachab, and A. Jangra, "A comprehensive survey of prominent cryptographic aspects for securing communication in post-quantum iot networks," *Internet of Things*, vol. 9, p. 100174, 2020.

[15] S. Hamad, K. Alheeti, Y. Ali, and S. Shaker, "Clustering and analysis of dynamic ad hoc network nodes movement based on fcm algorithm," 2020.

[16] H. Mohammad and A. C. Sastry, "Acnm: Advance coupling network model sleep/awake mechanism for wireless sensor networks," *International Journal of Engineering & Technology*, vol. 7, no. 1.1, pp. 350–354, 2018.

[17] R. R. Bhandari and K. Rajasekhar, "Mobility aware clustering routing algorithm (macron) to improve lifetime of wireless sensor network," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 2, p. 76, 2019.

[18] N. Srikanth and M. S. Prasad, "Energy efficient trust node based routing protocol (eetrp) to maximize the lifetime of wireless sensor networks in plateaus," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 15, no. 06, pp. 113–130, 2019.

[19] R. R. Bhandari and K. Rajasekhar, "Energy-efficient routing-based clustering approaches and sleep scheduling algorithm for network," *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2019*, vol. 89, p. 293, 2019.

[20] P. Goswami, Z. Yan, A. Mukherjee, L. Yang, S. Routray, and G. Palai, "An energy efficient clustering using firefly and hml for optical wireless sensor network," *Optik*, vol. 182, pp. 181 – 185, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0030402618320540

[21] K. R. Nirmal and K. Satyanarayana, "Map reduce based removing dependency on k and initial centroid selection mr-redic algorithm for clustering of mixed data," *measurement*, vol. 5, p. 6.

[22] H. Oh and C. T. Ngo, "A slotted sense multiple access protocol for timely and reliable data transmission in dynamic wireless sensor networks," *IEEE Sensors Journal*, vol. 18, no. 5, pp. 2184–2194, 2018.

[23] S. K. A. AYUSHREE, "Comparative analysis of aodv and dsdv using machine learning approach in manet," *Journal of Engineering Science and Technology*, vol. 12, no. 12, pp. 3315–3328, 2017.

Cluster-based Access Control Mechanism for Cellular D2D Communication Networks with Dense Device Deployment

Thanh-Dat Do[1], Ngoc-Tan Nguyen*[2], Thi-Huong-Giang Dang[3], Nam-Hoang Nguyen[4], Minh-Trien Pham[5]
VNU University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam[1,3,4,5]
Thang Long University, Hanoi, Vietnam[2]
University of Economics-Technique and Industry, Hanoi, Vietnam[3]

***Abstract*—In cellular device-to-device (D2D) communication networks, devices can communicate directly with each other without passing through base stations. Access control is an important function of radio resource management which aims to reduce frequency collision and mitigate interference between user's connections. In this paper, we propose a cluster-based access control (CBAC) mechanism for heterogeneous cellular D2D communication networks with dense device deployment where both the macro base station and smallcell base stations (SBSs) coexist. In the proposed CBAC mechanism, relied on monitoring interference from its neighboring SBSs, each SBS firstly selects their operating bandwidth parts. Then, it jointly allocates channels and assigns transmission power to smallcell user equipments (SUEs) for their uplink transmissions and users using D2D communications to mitigate their interference to uplink transmissions of macrocell user equipments (MUEs). Through computer simulations, numerical results show that the proposed CBAC mechanism can provide higher network throughput as well as user throughput than those of the network-assisted device-decided scheme proposed in the literature. Simulation results also show that SINR of uplink transmissions of MUEs and D2D communications managed by the MBS can be significantly improved.**

***Keywords*—D2D communications; access control; channel allocation; power assignment; interference mitigation**

I. Introduction

Future mobile networks are expected to provide communication services to billions of user equipments (UEs), i.e., regular mobile users and machine-type communication devices. In these networks, devices require a huge wireless traffic demand of device-to-device (D2D) communications such as vehicle to vehicle communications, communications between IoT devices. In traditional cellular networks, a base station (BS) acts as a relay to provide D2D communications for its users. Recently, cellular networks with D2D communications allow two arbitrary devices to directly establish a D2D communications link. With the aid of D2D communications, these networks can obtain significant improvements in terms of spectrum reuse, traffic offloading, low latency, and system throughput [1]-[4]. Nonetheless, cellular networks with D2D communications also bring lots of technical challenges such as high signaling load and frequency collisions (which may cause the degradation of the signal-to-interference-plus-noise ratio (SINR)).

*Corresponding Author: Ngoc-Tan Nguyen

In cellular networks with D2D communications, there are two type of communications, i.e., conventional cellular communications between BSs and their UEs, and D2D communications between two UEs. In the inband-overlay mode, different frequency bands are allocated for cellular and D2D communications, thus D2D communications cannot cause interference to cellular communications. Thus, the quality of service (QoS) of cellular communications is not affected by D2D communications but the efficiency of spectrum utilization is typically low [4]. By contrast, in the inband-underlay mode, a same frequency spectrum is allocated for both cellular and D2D communications. In this mode, the signals of a D2D communications might cause D2D-to-cellular interference to cellular communications when they use same channels [5]-[8], [9]. It is worth noting that cellular communications are given higher priority than D2D communications. To mitigate D2D-to-cellular interference, an efficient access control mechanism including channel allocation and transmission power assignment is needed to handle D2D connection requests.

Channel allocation, transmission power assignment, and interference mitigation are crucial research issues in cellular D2D communications networks. Power control can be implemented in different approaches, i.e., centralized manner or distributed manner [10]-[11]. In [10], the authors show that power control using the centralized algorithm can obtain higher performance than that using the distributed algorithm, but it suffers higher overhead as the number of devices increases. By contrast, in the distributed algorithm, D2D users exploit local channel state information to decide the transmission power, thus the overhead can be reduced. However, they might use high transmission power which can cause high interference to cellular users. Similar results for both centralized and distributed algorithms are also presented in the work [11]. The work in [12] provides a survey of radio resource management methods to reduce interference between D2D users and cellular users. Channel reuse technique using orthogonal frequency-division multiplexing is proposed in [13] where D2D and cellular users can share same spectrum. In the work [14], a joint admission control and resource allocation strategy is proposed to provide QoS support to cellular and D2D communications. The authors in [15] propose a resource scheduling method based on user location. However, in dense device networks, processing information of users' locations might cause high computational load. A guard zone based D2D-activation scheme is proposed in [16] in which the

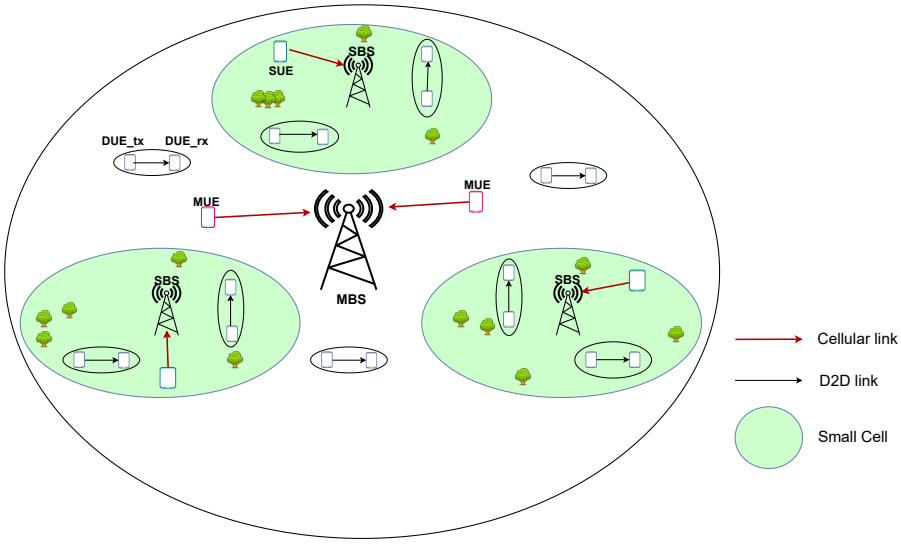


Fig. 1. A Dense Heterogeneous Cellular Network with D2D Communications.

exact closed-form expressions for the successful transmission probability of cellular users are proposed under the assumption that D2D users are uniform distributed within a geographical area. The scheme optimizes the guard zone's inner radius under the criteria of maximizing both transmission power and average throughput. Yet, this approach is only efficient for mobile networks with low user density.

The aforementioned works only study separated problems of channel allocation and power control which might lower the system performance. The authors in [9] analyze the impacts of co-channel interference between D2D links. Co-channel interference is unavoidable when the device density is ultra high. By joint optimizing channel allocation and power control, the interference mitigation efficiency and system performance can be significantly improved. A two-stage energy efficient maximization method including the power control and the channel allocation algorithms to improve D2D pair energy efficiency is proposed in [17] for D2D networks with low density. For dense D2D communication networks, the proposed method might cause high computation load. In [18], a distributed channel allocation and power control method based on Stackelberg game for D2D underlaid cellular networks is proposed to improve the sum-rate of D2D communications while meeting the QoS requirements of cellular users. This method can reduce the computation load of the base station effectively in small D2D communication networks. In [19], the authors propose a centralized resource management mechanism including channel allocation and transmission power control for heterogeneous cellular networks assisted by D2D communications. This mechanism can significantly improve the system throughput by mitigating D2D-to-cellular interference. However, it requires high computation load of the MBS and the channel measurement capability of UEs as the numbers of UEs and channels increase.

To our best knowledge, a practical access control mechanism for dense heterogeneous cellular networks with D2D communication assistance in which both the MBS and smallcell base stations (SBSs) coexist has not been investigated in the literature. In this paper, new constraints including the dense deployment of UEs and SBSs, flexible spectrum management (i.e., allocating multi bandwidth parts (BWPs) for UEs), and signaling load requirements are considered for the proposed heterogeneous cellular network assisted by D2D communications. Then, we propose a cluster-based access control mechanism involving the BWP selection for SBSs, channel allocation, and power assignment to smallcell user equipments and users using D2D communications to mitigate D2D-to-cellular interference as well as enhance network throughput.

The remainder of the paper is organized as follows. The system model of the proposed heterogeneous cellular D2D communication networks with dense device deployment is described in Section 2. Section 3 presents the proposed cluster-based access control mechanism. Simulation results and discussions are provided in Section 4. Finally, the conclusion is presented in Section 5.

II. System Model

A. System Model

As illustrated in Fig. 1, the proposed system model consists of a macro base station (MBS) and S smallcell base stations (SBSs) located randomly in the coverage area of the MBS to increase the network capacity. There are three considered types of user equipments (UEs): (1) macrocell UEs (MUEs) served by the MBS, (2) smallcell UEs (SUEs) served by the according SBS, (3) users using D2D communications (DUEs). Under this setting, D2D communications allow the DUEs to exchange their data to each other directly to provide low latency communications. The coverage area of MBS is divided to multiple sectors in which the uplink transmissions to the MBS and SBSs, and D2D communications share a same frequency spectrum of N_C channels. This spectrum is divided to M bandwidth parts (BWPs). Each uplink transmissions to the MBS and SBSs utilizes one channel belonging to a BWP. A D2D communication between a pair of DUEs is also allocated one channel of a BWP. Each BWP has N_C^{BWP} channels for data transmissions and one reference signal (RS) channel for

TABLE I. MATCHING TABLE BETWEEN THE SPECTRUM EFFICIENCY AND SINR [19]

Modulation	Code Rate (Default Repetition=1)	Spectrum Efficiency η (bps/Hz)	Minimum SINR (dB)
QPSK	$1/2(4)$	0.25	-2.5
QPSK	$1/2(2)$	0.5	0.5
QPSK	$1/2$	1	3.5
QPSK	$3/4$	1.5	6.5
16-QAM	$1/2$	2	9
16-QAM	$3/4$	3	12.5
64-QAM	$1/2$	3	14.5
64-QAM	$2/3$	4	16.5
64-QAM	$3/4$	4.5	18.5

broadcasting the RS. When a SBS uses a BWP, the SBS broadcasts the predefined RS on its RS channel of the BWP at a fixed transmission power. A SBS forms a D2D cluster and is allowed to use up to m BWPs of the total M BWPs of the network. Among of them, m_{SUE} BWPs can be used for uplink transmissions of SUEs and m_{D2D} BWPs can be allocated to D2D communications.

B. Pathloss Models and Interference Analysis

In the literature, various channel models are considered for the D2D communication networks [19]-[22]. In this paper, channel models proposed in [19] are adopted for the performance comparison. The channel model includes two transmission modes, i.e., Line-of-Sight (LOS) and non-Light-of-Sight (NLOS). The pathloss of these transmission modes are estimated as follows:

- The LOS pathloss model is applied to calculate the pathloss between the MBS and MUEs, the MBS and its DUEs, a SBS and its SUEs, and a SBS and its DUEs. The LOS pathloss is calculated as follows:

$$PL\left(d\right) = 127 + 30\log_{10}\left(d\right) + \varsigma. \tag{1}$$

- The NLOS pathloss model is applied to the D2D communications and uplink channels between DUEs and MUEs, DUEs and SUEs, and MUEs and SUEs. The NLOS pathloss is calculated as follows:

$$PL\left(d\right) = 128.1 + 37.6\log_{10}\left(d\right) + \varsigma, \tag{2}$$

  where d is the distance between a sender and a receiver in kilometers. ς is the shadowing in dB which follows log-normal distribution with the mean is zero and the standard deviation is one.

The total throughput obtained by a UE (i.e., a MUE, SUE, or DUE) is calculated as follows:

$$C = \sum_{i=1}^{N_{ch}} \eta_i B_i, \tag{3}$$

where N_{ch} is the number of channels used by the UEs. B_i denotes the bandwidth of the channel *i*-th among N_{ch} channels. The spectrum efficiency of the channel *i*-th which (denoted as η_i) depends on the SINR measured at the receiver (as shown in Table I [19]). To increase the spectrum efficiency of the proposed system, channels are reused in both D2D communications and uplink transmissions in the MBS and SBSs. However, this leads to the co-channel interference to the MUE. Specifically, when a SBS allocate a channel which is using by the MUE to its SUEs or DUEs, this can cause co-channel interference to the MUE. The D2D communications managed by the MBS can also interfere the uplink transmission of the MUE if they use same channel. Therefore, there is a need of an access control mechanism to allocate channel and optimize transmission power to minimize the co-channel interference.

III. CLUSTER-BASED ACCESS CONTROL MECHANISM

In order to mitigate co-channel interference as well as improve the system throghput, in this section, a cluster-based access control (CBAC) mechanism, which consists of channel allocation and power control, is designed for both the MBS and SBSs in the proposed D2D mobile network with the dense deployment of SBSs and DUEs. Firstly, a SBS selects its BWPs and forms its cluster (involving its SUEs and DUEs). The SBS can accept or remove a UE (i.e. a SUE or DUE) when the UE enters or moves out its coverage area, respectively. Then, the SBS performs the proposed CBAC mechanism to SUEs and DUEs those are in its cluster. While, the MBS performs the proposed CBAC mechanism to its MUEs and DUEs those do not belong to any cluster.

The functions of the MBS and SBSs are listed in detail as follows:

A. A SBS Selects BWPs

1) When a SBS configures its operating BWPs, it measures energy levels of RS channels of M BWPs, and then selects m BWPs having the lowest energy levels among M BWPs.
2) Then, among m selected BWPs, it assigns m_{D2D} BWPs for D2D communications which have lower RS energy levels than those of the remainder (m_{SUE}) BWPs for cellular communications.
3) Finally, it informs the MBS about its BWP configuration. The MBS is then responsible to update the number of SBSs using the same BWP.

B. A SBS Manages its Cluster and Estimate the Maximum Acceptable Interference

1) *Accept a new UE:* The SBS periodically broadcasts its pilot signal. If a new UE wants to be served by an

SBS, it detects a SBS with the strongest pilot signal among all SBSs in its range. Then, it sends a request to that SBS for cluster registration. If the SBS still has an available room for the UE, it accepts the UE to join its cluster. Otherwise, the SBS rejects and informs the UE to look for another available SBS.

2) *Remove an inactive UE:* A UE might be inactive when it is off or moves out the coverage area of its serving SBS. The serving SBS periodically asks its UEs to confirm whether they are still active or not. If the serving SBS does not receive any confirmation from a UE, the serving SBS can remove the UE out of its cluster.
3) *The SBS estimates the maximum acceptable interference:* It is worth noting that the channel allocated to a D2D communication or an uplink cellular transmission to the SBS might be also used by a MUE. In that case, the SUEs or DUEs (served by the SBS and MBS) can cause interference to the MUE. Therefore, the SBS must control the transmission power of its SUE and source DUE subject to their interference to the MUE that does not exceed the maximum acceptance interference. It is worth noting that the worst case of an edge MUE which is most vulnerable to interference from SUEs and DUEs is considered. Firstly, the MBS informs the SBS about the maximum acceptable interference I_{total}^{max} that the edge MUE still can guarantee its SINR threshold:

$$I_{total}^{\max} = \frac{P_{max}^{MUE}}{\gamma^0 PL\left(R\right)}, \tag{4}$$

where P_{max}^{MUE} is the maximum transmission power of the MUE. $PL(R)$ is the estimated pathloss of the channel link from the edge MUE to the MBS with R is the radius of the MBS's cell. γ^0 denotes the SINR threshold of the uplink transmission from the MUE to the MBS. Under the worst setting, all neighbors of the SBS also allocate the same channel to their SUEs or DUEs. Assume that the SBS can detect N_{SBS} neighbors, then the SBS can calculate the maximum acceptable interference that its DUE or SUE can cause to the edge MUE:

$$I_{SBS}^{max} = \frac{I_{total}^{\max}}{N_{SBS}+1}. \tag{5}$$

C. A SBS Allocates Channel And Assign Transmission Power To Its SUEs

1) When a SUE wants to establish an uplink connection with its SBS, the SUE firstly sends a connection request to its SBS.
2) Then, the serving SBS finds a BWP that has available channels and the lowest RS energy level among m_{SUE} BWPs. If a qualified BWP is found, the SBS allocates a round-robin free channel of the BWP to the SUE with an uplink transmission power P_{SBS}^{SUE} assigned as follows:

$$P_{SBS}^{SUE} = I_{SBS}^{max} PL\left(d_{MBS\to SBS}\right), \tag{6}$$

where $PL\left(d_{MBS\to SBS}\right)$ is the estimated pathloss from the MBS to the serving SBS. Assume that the pathloss information of SUEs is not available at the serving SBS (since it does not know the exact locations of SUEs). Thus, the pathloss from SUEs to the SBS is approximateed as the pathloss from the MBS to SBS.

D. A SBS Allocate Channel And Assign Transmission Power To Its DUEs

1) When a DUE wants to establish a D2D communication with another DUE, the DUE firstly sends a D2D connection request to its serving SBS.
2) Then, the serving SBS finds a BWP which has available channels and the lowest RS energy level among m_{D2D} BWPs. If a qualified BWP is found, the SBS allocates a round-robin free channel of the BWP to the DUE with a transmission power P_{SBS}^{DUE} assigned as follows:

$$P_{SBS}^{DUE} = I_{SBS}^{max} PL\left(d_{MBS\to SBS}\right). \tag{7}$$

E. The MBS Assigns Channel and Transmission Power to its MUEs

1) When a MUE wants to establish an uplink transmission to the MBS, the MUE firstly sends a connection request to the MBS.
2) Then, the MBS allocates a round-robin free channel to the MUE and assign an initial transmission power P_{MBS}^{MUE} to the allocated channel.
3) After the uplink transmission between the MUE and the MBS is established, they collaborate to optimize the uplink transmission power for the MUE. Firstly, the MBS measures the SINR of the uplink channel and finds an optimal transmission power (that guarantees the SINR target) for the MUE. Then, it sends a power control message with the optimal transmission power information to the MUE.

F. The MBS Assigns a Available Channel and Transmission Power to DUEs that do not Belong to any Clusters

1) When a DUE, which is currently served by the MBS, wants to establish a D2D communication with another DUE, it sends a D2D connection request to the MBS. Then, the MBS allocates a round-robin free channel of its BWP to the DUE.
2) The worst case is considered in which all SBSs also allocate the same channel to its UEs that causes interference to the D2D communications served by the MBS. Since the MBS knows the number of SBSs (denoted as $N_{SBS}^{'}$) which are using the same BWP. Then, the MBS estimates the maximum acceptable interference (I_{MBS}^{max}) that the SBS's UEs and the DUE (served by the MBS) can cause to the MUE:

$$I_{MBS}^{max} = \frac{I_{total}^{\max}}{N_{SBS}^{'}+1} \tag{8}$$

3) Finally, the MBS assigns the transmission power to the DUE:

$$P_{MBS}^{DUE} = I_{MBS}^{max} PL\left(d_{DUE\to MBS}\right), \tag{9}$$

where $PL\left(d_{DUE\to MBS}\right)$ is the estimated pathloss from the DUE to the MBS.

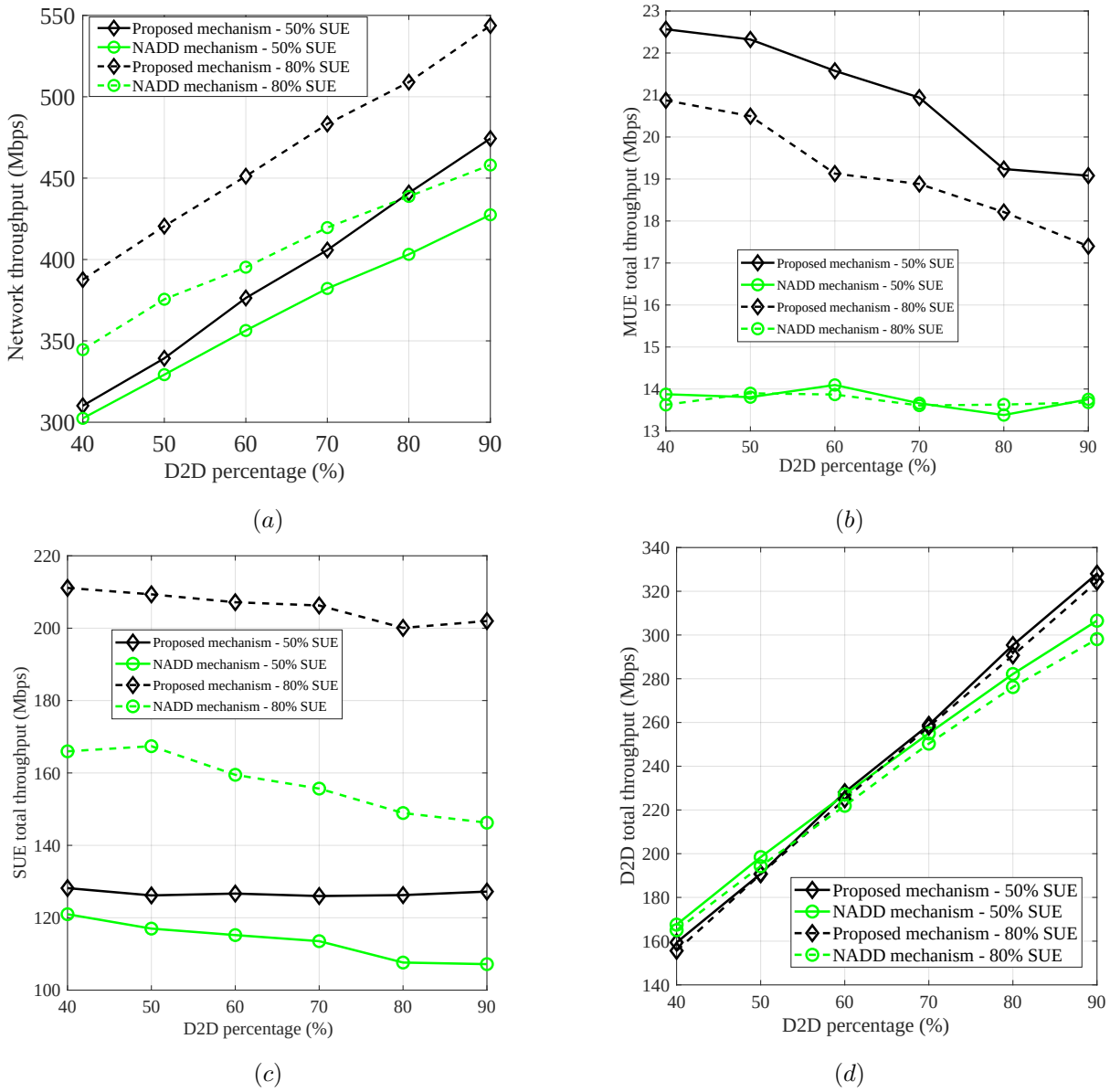


Fig. 2. (a) Network Throughput, (b) Total Throughput Obtained by MUEs, (c) Total Throughput Obtained by SUEs, (d) Total Throughput Obtained by D2D Communications vs. D2D Percentage.

IV. Simulations and Numerical Results

In the section, we conduct computer simulations and performance evaluations of the proposed cluster-based access control (CBAC) mechanism and the dynamic network assisted device decided (NADD) mechanism proposed in [19]. In our simulations, we consider a MBS and 50 SBSs under the coverage of the MBS (having a radius of 1000m). Each SBS has the coverage radius of 100m. The spectrum has 60 channels divided into 6 BWPs and each channel has the bandwidth of 180 KHz. Under this setting, each MUE or SUE uses one channel for their uplink transmissions and the other channel is used for D2D communications. The maximum transmission power of MUEs, SUEs, and DUEs is 23 dBm [23]-[25]. Each SBS is assumed to consume two BWPs, the former is used for SUEs' uplink transmissions and the latter is used for D2D communications. The SUE and D2D percentages of a SBS are defined as the ratios of the number of simultaneous SUEs' uplink transmissions and D2D communications to the total number of available channels of the SBS, respectively. Other setting parameters are listed in Table II.

A. Throughput Performance

For throughput performance comparison, two scenarios, i.e., different D2D percentages and SINR thresholds, are investigated to evaluate the throughput performance.

1) Varying D2D Percentage: Fig. 2(a) - Fig. 2(d) show the comparisons of the network throughput, total throughput obtained by MUEs, SUEs, and DUEs (via D2D communications), respectively, when varying the D2D percentage. Two different

TABLE II. SIMULATION PARAMETERS.

Setting Parameters	Value	Unit
Macrocell radius (R)	1000	m
Number of channels	60	channel
Number of BWPs	6	BWP
Number of SBSs	50	SBS
Number of BWPs in a SBS	2	BWP
Smallcell radius	100	m
Bandwidth of a subchannel	180	KHz
Number of macrocell UEs	50	MUE
Number of SUEs in a SBS	8	SUE
MUE channel usage	1	channel
SUE channel usage	1	channel
D2D channel usage	1	channel
Maximum transmission power of MUE/device	23	dBm
Mean distance between two devices in D2D pair	20	m
MUE's SINR target for power control	20	dB
MUE's SINR threshold	7.5	dB
Carrier frequency	2.0	GHz

settings of the SUE percentage in each SBS (i.e., 50% and 80%) have been considered in all simulations. In Fig. 2(a), simulation results show that the proposed CBAC mechanism provides higher network throughput than that of the NADD mechanism. For example, when the SUE and D2D percentage are 50% and 70%, respectively, the network throughput of the proposed CBAC mechanism is 10% higher than that of the NADD mechanism. As the D2D traffic load increases, the network throughput also increases since SBSs can accept any new D2D connection requests until all D2D channels of SBSs are occupied. However, when the D2D traffic load increases, the throughput obtained by MUEs and SUEs are decreased as shown in Fig. 2(b) and Fig. 2(c), respectively. It is due to the fact that D2D communications can cause interference to MUEs and SUEs. Thus, the more D2D communications, the higher interference to MUEs and SUEs. It is recommended that in the cellular D2D mobile network, it is necessary to set a limit on the number of D2D communications. As can be seen in Fig. 2(d), the total throughput obtained by D2D communications of the proposed CBAC mechanism is higher than that of the NADD mechanism. The reason is that in the proposed CBAC mechanism, a SBS can select BWPs with low interference levels for D2D communications which results in lower interference from other SBSs and MUEs.

2) Varying SINR threshold: Fig. 3 illustrates the total throughput obtained by MUEs, SUEs and D2D communications as varying the SINR threshold of the MUE under the setting of the SUE and D2D percentages at 80%. When the SINR threshold of the MUE increases, the maximum acceptable interference is decreased which results in reducing the transmission power of SUEs and DUEs. Therefore, as shown in Fig. 3, the total throughput obtained by MUEs is increased. Overall, the proposed CBAC mechanism has higher total throughput obtained by MUEs, SUEs and D2D communications than those of the NADD mechanism. The reason is that the proposed CBAC mechanism is able to avoid the frequency collision between uplink transmissions of SUEs and D2D communications, thus reduce the co-channel interference from SUEs and D2D communications to MUEs.

B. Interference Mitigation

To evaluate the interference mitigation of the proposed CBAC mechanism for co-channel interference from uplink

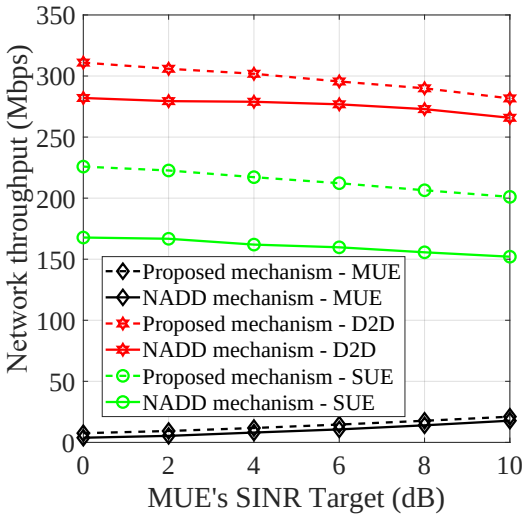


Fig. 3. Network Throughput vs MUE's SINR Threshold.

transmissions of SUEs and D2D communications to MUEs, we plot the statistical cumulative distribution functions (CDFs) w.r.t. SINRs of received signals of the MUEs' and SUEs' uplink transmissions, and D2D communications, respectively, which are measured at the MBS when the SUE percentage of each SBS is 80%. As shown in Fig. 4(a), at the SINR value of 7dB, the NADD mechanism can provide 30% of SINR samples less than 7dB whereas the proposed CBAC mechanism provides 40% of SINR samples less than 7dB. However, at the higher SINR value, e.g., 10dB, the proposed and NADD mechanisms provide about 45% and 90% of SINR samples less than 10dB, respectively. That means the proposed CBAC mechanism is able to mitigate the interference from SUEs and D2D communications to MUEs in the case of dense device deployment. Fig. 4(b) and Fig. 4(c) also show that the proposed CBAC mechanism provides better SINR than the NADD mechanism. It is due to the fact that in the proposed CBAC mechanism, each SBS forms a cluster and allocates different BWPs to SUEs' uplink transmissions and D2D communications in its cluster, thus the interference between SUEs' uplink transmissions and D2D communications are locally eliminated which results in the improvement of SINR of SUEs and D2D communications.

V. CONCLUSION

In this paper, we have studied a heterogeneous cellular D2D communication networks with new constraints of dense device deployment, flexible spectrum management and low signaling load requirements. We have then proposed the cluster-based access control (CBAC) mechanism to mitigate D2D-to-cellular interference and enhance network throughput. Specifically, in the proposed mechanism, each SBS firstly forms a cluster of SUEs and DUEs, and selects qualified BWPs. Then, it jointly performs channel allocation and transmission power assignment to its SUEs or DUEs based on the estimated maximum D2D-to-cellular interference. Simulation results have proved that the proposed CBAC mechanism can provide higher network throughput as well as total throughput obtained by MUEs, SUEs, and DUEs (via D2D communications). There

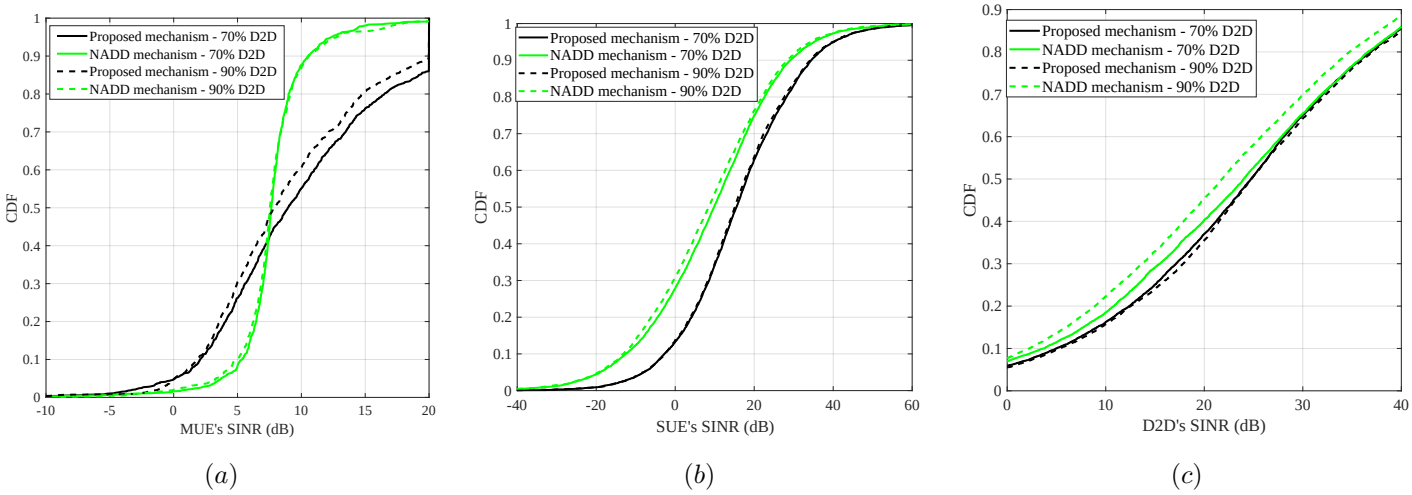


Fig. 4. (a) CDF of MUE's SINR, (b) CDF of SUE's SINR, (c) CDF of D2D's SINR.

are still open research issues for future research in resource management for heterogeneous cellular D2D communication networks with dense device deployment such as the cooperative access control between the MBS and SBSs, or a distributed transmission power optimization and interference mitigation problem.

ACKNOWLEDGMENT

This work has been supported by Vietnam National University, Hanoi (VNU), under Project No. QG.19.24.

REFERENCES

[1] W. Cao, G. Feng, S. Qin and M. Yan, "Cellular Offloading in Heterogeneous Mobile Networks With D2D Communication Assistance", *IEEE Transactions on Vehicular Technology,* vol. 66, pp. 4245-4255, Aug. 2016.

[2] Y. Niu et al., "Exploiting Device-to-Device Communications to Enhance Spatial Reuse for Popular Content Downloading in Directional mmWave Small Cells", *IEEE Transaction on Vehicular Technology,* vol. 65, pp. 5538-5550, Aug. 2015.

[3] W. Lee, J. Kim, and S. Choi, "New D2D Peer Discovery Scheme based on Spatial Correlation of Wireless Channel", *IEEE Transaction on Vehicular Technology,* vol. 66, pp. 10120–10125, Feb. 2016.

[4] M. Hicham, N. Abghour, and M. Ouzzif, "Device-To-Device (D2D) Communication Under LTE-Advanced Networks", *International Journal of Wireless & Mobile Networks,* vol. 8, pp. 11-22, Feb. 2016.

[5] J. Huang et al., "Modeling and Analysis on Access Control for Device-to-Device Communications in Cellular Network: A Network Calculus Based Approach", *IEEE Transactions on Vehicular Technology,* vol. 65, pp. 1615–1626, Mar. 2016.

[6] P. Mach, Z. Becvar, and T. Vanek, "In-band Device-to-Device Communication in OFDMA Cellular Networks: A Survey and Challenges", *IEEE Communications Surveys & Tutorials,* vol. 17, pp. 1885-1922, Jun. 2015.

[7] C. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks," *IEEE Transactions on Wireless Communications,* vol. 10, no. 8, pp. 2752-2763, Aug. 2011.

[8] A. Bhardwaj and S. Agnihotri,, "Energy- and Spectral-Efficiency Trade-Off for D2D-Multicasts in Underlay Cellular Networks", *IEEE Wireless Communications Letters,* vol. 7, pp. 546–549, Aug. 2018.

[9] G. A. Safdar, M. Ur-Rehman, M. Muhammad, M. A. Imran, and R. Tafazolli, "Interference Mitigation in D2D Communication Underlaying LTE-A Network", *IEEE Access*, vol. 4, pp. 7967–7987, Oct. 2016.

[10] N. Lee, X. Lin, J. G. Andrews, and R. W. Heath, "Power Control for D2D Underlaid Cellular Networks: Modeling, Algorithms and Analysis", *IEEE Journal on Selected Areas in Communications,* vol. 33, pp. 1–13, Jan. 2015.

[11] W. Lee, T. Ban and B. C. Jung, "Distributed Transmit Power Optimization for Device-to-Device Communications Underlying Cellular Networks", *IEEE Access,* vol. 7, pp. 87617–87633, Jul. 2019.

[12] P. Gandotra and R. K. Jha, "Device-to-Device Communication in Cellular Networks: A Survey", *Journal of Network and Computer Applications,* vol. 71, pp. 99-117, Aug. 2016.

[13] F. Berggren and B. M. Popović, "Primary Synchronization Signal for D2D Communications in LTE-Advanced", *IEEE Communications Letters,* vol. 19, pp. 1241-1244, Jul. 2015.

[14] S. Cicalò and V. Tralli, "QoS-Aware Admission Control and Resource Allocation for D2D Communications Underlaying Cellular Networks", *IEEE Transactions on Wireless Communications,* vol. 17, pp. 5256-5269, Aug. 2018.

[15] D. Feng et al., "Device-to-Device Communications in Cellular Networks", *IEEE Communications Magazine,* vol. 52, pp. 49–55, May 2014.

[16] S. Lv, C. Xing, Z. Zhang and K. Long, "Guard Zone Based Interference Management for D2D-Aided Underlaying Cellular Networks", *IEEE Transactions on Vehicular Technology,* vol. 66, pp. 5466–5471, Oct. 2016.

[17] S. Liu, Y. Wu, L. Li, X. Liu and W. Xu, "A Two-stage Energy-efficient Approach for Joint Power Control and Channel Allocation in D2D Communications", *IEEE Access,* vol: 7, pp: 16940–16951, Jan. 2019.

[18] Y. Yuan, T. Yang, H. Feng and B. Hu, "An Iterative Matching-Stackelberg Game Model for Channel-power Allocation in D2D Underlaid Cellular Networks", *IEEE Transactions on Wireless Communications,* vol: 7, pp: 7456-7471, Sep. 2018.

[19] S. Yang, L. Wang, J. Huang and A. Tsai, "Network-assisted Device-decided Channel Selection and Power Control for Multi-pair Device-to-Device (D2D) Communications in Heterogeneous Networks", in *IEEE Wireless Communications and Networking Conference,* Nov. 2014.

[20] B. Zhou, H. Hu, S. Huang and H. Chen, "Intracluster Device-to-Device Relay Algorithm with Optimal Resource Utilization", *IEEE Transactions on Vehicular Technology,* vol. 62, pp. 2315-2326, Jan. 2013.

[21] J. Ding, L. Jiang and C. He, "Energy-Efficient Power Control for Underlaying D2D Communication with Channel Uncertainty: User-Centric versus Network-Centric", *Journal of Communications and Networks,* vol. 18, pp. 589–599, Aug. 2016.

[22] H. Xu, W. Xu, Z. Yang, J. Shi and M. Chen, "Pilot Reuse among D2D Users in D2D Underlaid Massive MIMO Systems", *IEEE Transactions on Vehicular Technology,* vol. 67, pp. 467-482, Jul. 2017.

[23] H. Sun, M. Wildemeersch, M. Sheng and T. Q. S. Quek, "D2D Enhanced Heterogeneous Cellular Networks with Dynamic TDD", *IEEE Transactions on Wireless Communications,* vol. 14, pp. 4204-4218, Mar. 2015.

[24] H. Tang, Z. Ding and B. C. Levy, "Enabling D2D Communications through Neighbor Discovery in LTE Cellular Networks", *IEEE Transactions on Signal Processing,* vol.62, pp. 5157-5170, Aug. 2014.

[25] G. Liu, W. Feng, Z. Han and W. Jiang, "Performance Analysis and Optimization of Cooperative Full-Duplex D2D Communication Underlaying Cellular Networks", *IEEE Transactions on Wireless Communications,* vol.18, pp. 5113-5127, Aug. 2019.

Human Recognition using Single-Input-Single-Output Channel Model and Support Vector Machines

Sameer Ahmad Bhat[1], Abolfazl Mehbodniya[2], Ahmed Elsayed Alwakeel[3], Julian Webber[4] and Khalid Al-Begain[5]

Dept. of Electronics and Communications Engineering, Kuwait College of Science and Technology (KCST), Kuwait[1,2,3,5].
Graduate School of Engineering Science, Osaka University, Japan[4].

***Abstract*—WiFi based human motion recognition systems mainly rely on the availability of Channel State Information (CSI). Embedded within WiFi devices, the present radio subsystems can output CSI that describes the response of a wireless communication channel. Radio subsystems as such, use complex hardware architectures that consume lots of energy during data transmission, as well as exhibit phase drift in the sub-carriers. Although human motion recognition (HMR) based on multi-carrier transmission systems show better classification accuracy, transmission of multiple sub-carriers results in an increase in the overall energy consumption at the transmitter. Apparently CSI based systems can be perceived as process intensive and power hungry devices. To alleviate the process intensive computing and reduce energy consumption in WiFi, this study proposes a human recognition system that uses only one radio carrier frequency. The study uses two software defined radios and a machine learning classifier to identify four humans, and the study results show that human identification is possible with 99% accuracy using only one radio carrier. The results of this study will have an impact on the development process of smart sensing systems, particularly those that relate to healthcare, authentication, and passive monitoring and sensing.**

Keywords*—Motion detection; pattern recognition; received signal strength indicator; Software Defined Radio (SDR); supervised learning

I. Introduction

In recent years, the role of smart environments has attracted most of the research communities across the globe, and the research activities undertaken by such communities, are transforming the existing natural, or made-man setups to smart environments. The areas that are influenced by these transformations also include indoor sensing, pattern recognition and classification systems, and smart environments. Smart home applications, spanning across various domains, enable support to build smart home environments, and human motion sensing environments, in particular, enable support to motion sensing, analysis, and evaluation of ambient environmental settings, or parameters, as a response to human activities.

At present, various state-of-the-art analytical methods have been devised to explore the analogous, and discriminative physiological and behavioral characteristics of humans, so as to model human motion behavior as well as to recognize different human motion patterns. A smart home is realized as a subunit in a smart environment, wherein human motion recognition, and localization applications may be deployed, both in indoor and outdoor environments. Indeed device free passive indoor localization [1] has been of great interest to researchers, and it plays a key role in the applications that enable assisted living facilities for elderly, children, physically challenged, and in smart home, etc. [2], [3], [4].

The combined motion sensing approaches and their reasoning deliver context-aware data from human motion, as well as from the analysis of human activities. The data collected is then next employed to provide personalized support in many applications [5]. Human motion sensing, analysis and prediction is classified into three categories: vision-based systems, wearable sensor-based systems, and RF based systems [6][7]. Sensing systems based on the approach of vision sensing are classified under the category of passive sensing systems. Sensing systems as such, use cameras as a light sensor for tracking human motion patterns [8] [9]. Human behavioral patterns captured in images can be processed using the computer vision and machine learning techniques. Typically, images captured with a camera, often needs a camera to have sufficient ambient light, and insufficient lighting effects visibility, which leads to significant decrease, or even no sensing capabilities in cameras. Moreover, any physical barriers such as walls completely alienate camera based sensing systems.

Wearable sensor-based approach [10] [11] [12], is one of the alternatives to monitoring human activities. Wearable sensing needs attaching sensors to the human body, and often pose challenges of electrical wiring, power supply management and in particular mainly cause inconvenience. Consequently, elderly patients abstain to carry electrical wires and monitoring sensors. Typically, subjects' data are recorded with inertial sensors such as gyroscopes, accelerometers, or magnetometers, enabling human motion data acquisition as electrical signals varying over time [13]. For consistent observation and monitoring of subjects' necessitates subjects carrying monitoring devices, often demanding power supply and other accessories items to supply. Thus, proposing solutions to eliminate sensor deployment on the human body is imperative and directs research to incorporate passive sensing and monitoring elements.

The shadowing effects left over by any moving targets intercepting line of sight (LoS) path between the transmitter and receiver, enable tracking of objects in motion in indoor environments [14] [15] [16]. An interception caused by a human walking across the LOS of the RF signal, results in variations in the received strength signal (RSS) at the receiver [17]. To Identify and track the unique human motion patterns out of RSS, requires analyzing the embedded unique human motion signatures, using various methods of signal processing.

On the other hand, advancements in wireless technology are driving researchers to devise solutions exploiting wireless communication systems in localization and pattern recogni-

tion based applications. In fact, several attempts addressing issues concerning motion detection [18][19], gesture detection [20][21], and facial recognition systems [22] have been successful. With current wireless devices embedding multiple radio sub-units, allow sub-carriers for data communications. However, SDRs can also provide estimated Channel State Information (CSI), and many commercially available off-the-shelf (COTS) devices support CSI data directly via in-built subsystems.

Although CSI based localization and pattern recognition studies reveal real higher performances, developed systems still exhibit challenges such as increased processing complexity, portability, adaptability, unreliability, lower precision, and inefficient system designs. Solutions based on existing wireless communication systems infrastructure, no doubt extend the scope of research in the current context, however, processing multiple sub-carriers in CSI based systems is of concern. Statistical CSI data retrieved either from commercially or customized firmware modified routers provides human activity and gesture data [23]. Recent attempts made using CSI of WiFi devices have shown higher motion recognition accuracies, nonetheless, it solely relies on available WiFi channels for monitoring. Therefore, developing efficient system designs will significantly impact potential applications implementing elementary and straightforward prototyping methods of localization and pattern recognition. This study aims to address the problem of human recognition. Our proposed solution bases on one sub-carrier frequency only, rather than multiple subcarrier frequencies, to identify and classify human motion patterns using machine learning.

The study will impact future works that relate to human recognition systems, employing SISO channel model of communication instead of using CSI based systems. The proposed testbed in this study can be used for further investigation of works like, random human motion detection, motion speed detection, trespasser detection, and many other applications, wherein human motion detection may be carried out passively from a remote monitoring station. Moreover, when using current system with multiple deployments at different locations, would eventually lead to a passive sensing system, analogous to sensor networks that transmits sensed information via nodes mounted at various location within a network. Thus, human recognition is possible under the domain of a passive sensing system.

A. Research Contribution

The main contributions of this paper are listed below:

- The study proposes a testbed for recognizing humans in indoor environments using two NI-USRPs, and it highlights the main challenges, experienced while setting up the testbed for the study.
- The study identifies possible setup for experiment, parameters that help in tuning of SDRs to the optimum level, along with setup to conduct further research in the domain of human recognition systems that may employ just single-input-single-output model of communication channel.
- The study provides a comparative analysis of two different machine learning models employed in this study for human identification. Moreover, the study accesses the level of accuracy of two different machine learning models that show an accuracy of 99% in identifying humans based on their patterns of locomotion.

The study is organized according to the following sections. Section II, provides a detailed background to the study. Section III provides a theoretical perspective, system model, and the method of data collection. Section IV describes the SVM based machine learning solution for pattern detection and classification. Section V provides a discussion on the study results and comparison of SVM performances. Finally, Section VI concludes the overall study.

II. Background

Wireless signal propagation is influenced by various environmental factors, wherein wireless signal strength is mainly attenuated by multipath fading, path loss, and shadowing. Multipath results in a transmitted signal to arrive at the receiver, as multiple reflections of the original signal, from different paths, thereby causing severe distortion in the original signal component. Whereas the signal strength attenuates due to increased propagation distance and mainly relies on channel behavior, shadowing results in power loss due to physical objects appearing in signal propagation path.

The CSI of subcarrier [23], [24] frequencies show random variation, with an added distortion as a result of reflections in multipath propagation of a signal. Typically, random variations observed under normal conditions describe dynamics of CSI, whilst without the presence of nearby objects within the range. The CSI pertaining to various sub-carrier frequencies require methods of signal processing to de-noise and decompose the distinguishing features embedded in the signal, should be extracted for the purpose of motion pattern recognition. On the other hand, random human motion inevitably influences CSI elementary behaviour, and extracting meaningful information even becomes harder. Prior studies conducted explore the dynamic nature of RSS and extract patterns by applying machine learning algorithms on acquired data from multiple sub-carriers. For example, [23], [24], [25] have applied Principal Component Analysis (PCA), [26] used Discrete Wavelet Transform (DWT),the CLEAN algorithm by [27], Doppler spectrum by [28] and even scale and time shift projections by [26], were used. Nonetheless, proposed signal processing and patterns recognition methods require high speed processing elements including scaled hardware resources that, in general, contribute to inefficient, unreliable and expensive methods of human recognition.

The random phase drift, as a result of sampling time offsets, is common in CSI phase measurements [27]. Consequently, for observed motion patterns, contrasting results can be seen with similar devices enabling CSI data generation. Processing devices with lower CSI-subcarrier sampling rate leads to processing delays and hence limits CSI based motion sensing in real time applications. Indoor environments include surrounding objects and motion observations may include background clutters. The Background elimination algorithm, for example by [27], subtracts static paths from the observed data, thus enables background clutter removal. While the likelihood criterion removes target reflection path from observed motion path, challenges in defining descriptor variables in noisy measurements still post serious concerns [28].CSI sub-carriers in turn reflect random noise intensity levels. Each CSI sub-carrier component requires processing at an individual level, thus adding requirements of additional processing.

MIMO systems [25] [27], on the other hand, employ multiple antennas in accretion to multiple receivers. Motion sensing systems based on MIMO, thus add increased device accessories and deployment costs. In addition, 5GHz CSI receivers implementing 114 and 132 sub-carriers for human motion sensing reveal only 94.0% accuracy [25], [24], meaning recognition accuracy is independent of added CSI channels. Here True Positive Rate (TPR) of CSI sub-carriers is independent, hence, added channels do not contribute to sensed motion recognition accuracy level.

As outlined in the previous sections, motion recognition based on CSI acquired from WiFi, is process intensive, and requires a lot of resources for processing than processing only one radio carrier frequency. Therefore, this study proposes a single-input-single-output (SISO) communication model based human motion recognition, and the proposed system is evaluated using testbed setup employing two National Instruments universal radio peripherals (NI-USRPs) to discover patterns embedded in the dynamics of a radio signal. The optimal configuration settings of TX / RX subsystem are highlighted, along with the method of experimentation, and how to apply AI for classifying different human motion patterns to reveal identities of people, in particular four participants. The proposed system aims to show that human recognition is possible with only one radio carrier, which is far better than CSI based human recognition system that reveals human motion signatures based on multiple radio subcarrier frequencies.

III. Methodology

Our experimental setup was based on two NI-USRP [29] SDR devices. SDRs alleviate the hardware and software level tuning, and initial experimental setup was based on NI-USRP 2901 model and LabVIEW Communication Design Software [30]. For initial systematic trials, randomized control trials (RCT) based setup was used to search for the optimal alignment parameters as well as control configuration of the used SDRs. For data collection, four participants were employed to walk through a predefined path, whilst following the directions given prior to the conduct of trails. Using the developed software application, enabled collection of RSSI patterns that embedded human motion signatures, in the spreadsheet files having CSV format. Our experimental foundation relies on the following theoretical perspective, provides bases to software application development to capture human motion data.

A. Transmitter-Receiver Sub-System

Our testbed setup implements a Continuous Wave (CW) transmitter modulated with 10kHz sine-tone. Both the transmitter and receiver sub-system are equipped with one omni-directional antenna (VERT2450) for transmission and reception respectively. For indoor environments, exhibit radio signal propagation characteristics wherein a transmitted signal converts to alias forms due to multipath propagation. In the current setup, transmitted CW arrives through multiple paths at receiver, and each different path adds delay and attenuation. CW transmission over longer distances, attenuates signal strength and wavelength considerably. In addition, RSS drastically varies due to small variations in multipath propagation, and variations equivalent to 5 dB in 1 minute have been observed in fixed receiver-transmitter pairs [31]. This study employs a fixed NI-USRP 2901 transmitter and receiver for effective results.

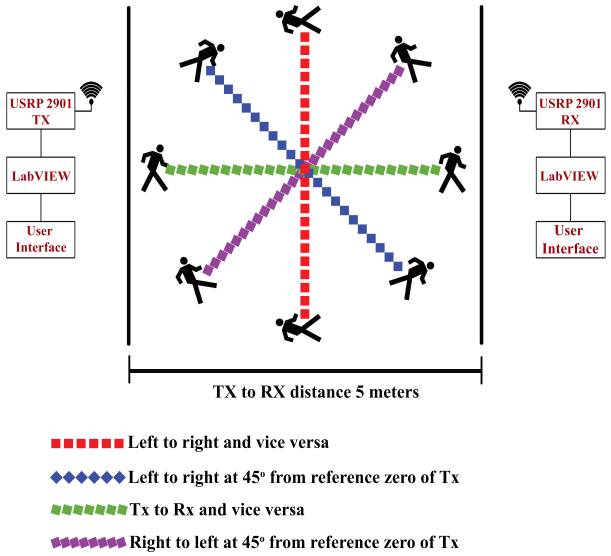


Fig. 1. Proposed System Model for Human Motion Patterns Recognition System.

CW signal strength arriving via mulipath at the receiver is given by the following relation:

$$V = \sum_{k=1}^{N} \|V_k\| \, e^{-j\theta_k} \tag{1}$$

where V_k, θ_k are the magnitude and phase of the kth multipath component respectively. Symbol N denotes total multipath components arriving at the receiver.

The NI USRP-2901 SDR model amplifies, down-converts, filters, digitizes, and decimates the received signal before the signal is transferred to the host computer. Similarly, the device up-samples, reconstructs, filters, up-converts, and amplifies the CW signal before its transmission into space. Testbed setup used two NI-USRP 2901 devices connected to two host computers. Fast Fourier Transform (FFT) is applied to extract frequency pilot signal in frequency domain. The demodulation process on the receiver recovers pilot signal strength, and the FFT provides the pilot RSS, which is expressed in decibel milliwatt (dBm) and given by relation:

$$\mathrm{RSS(dBm)} = 10\log_{10}\left(\|V\|^2\right) \tag{2}$$

B. System Model

Our proposed system model is based on two NI-USRP devices and required a software application for data collection and processing. Two NI-USRP devices were set up with a LOS separation distance of 5 meters. Both devices were placed on computer tables with a height 1 meter above the ground. NI-USRPs devices are connected to host computers via USB 3.0 cabling, and custom developed LabVIEW application software enabled data collection. As depicted in Fig. 1, our proposed system model implements single input single output (SISO) channel. Clutters in the background are removed in this experiment, and the lab environment is completely an open space holding two tables placed opposite to each other,

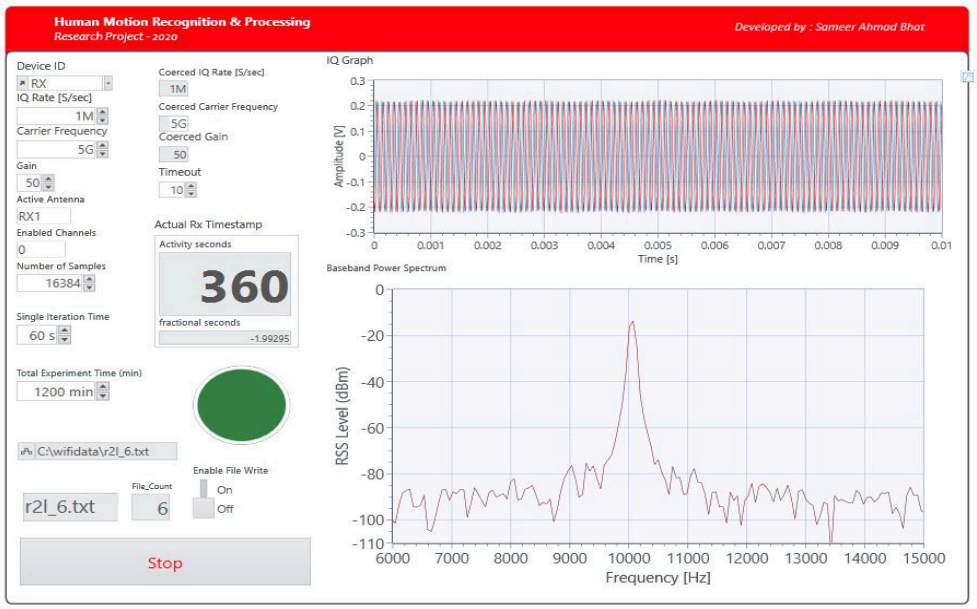


Fig. 2. Human Motion Recognition and Processing Application (HMRPA).

though a projector device is available in the room displaying the signal waveform on the side wall. Testbed setup using NI a single transmitter (TX) and receiver (RX) operate at 5 GHz band. The proposed system model allows monitoring motion (walking) patterns in four different directions. Testbed setup was complete in the two phases – Software Application setup and Software application development,and both phases were sequentially carried out:

1) Phase I – Application Development: A human motion recognition and processing application (HMRPA) (see Fig. 2) was developed. Nex the developed software module was added to the NI-USRP RX application. Initially, the preliminary study testbed supported measurements and recordings of motion patterns for a single participant only. The HMPRA software module was integrated with the modified NI receiver application available in LabVIEW communication design software. HMPRA allowed parameter setting like experiment time, sampling, walking interval spacing and also the signal processing logic at baseband level. The HMRPA generated raw text files containing RSS samples of executed motion in one direction only. Subsequently, the Randomized Control Trial (RCT) experiment allowed collecting multiple motion patterns of a single person in four different directions (see Fig.1).

2) Phase II – Hardware Setup: Two NI-USRP 2901 SDRs were employed to set up the testbed. Both TX and RX applications were custom developed in LabVIEW IDE (Communications Design Software CDS)). NI-USRP set in transmitter mode, transmit a pilot tone of 10KHz at a carrier frequency of 5GHz using an omni-directional antenna. At the receiver, the RX application implemented a Fast Fourier Transform (FFT), with the Power Spectral Density (PSD) enabled estimations of recovered side-tone. The baseband Power Spectral Density (PSD) of RSS provided best possible TX – RX orientation, whilst having no obstacles in the line of sight (LoS).

With no motion LoS path, RSS stayed showed small deviations, even when a person staying 10 meters away from LoS. However, any participant walking through the LoS pilot tone drastically changed the pilot side-tone magnitude. Variations as such contained motion patterns samples, which were output in text files by the software application. Prior to testing, TX and RX were properly aligned to show maximum and stable pilot side-tone signal strength, and optimum results were observed with individual gains of both set to 50. RSS power levels observations at different gain settings for TX and RX (see Table II), and RCTs for control group RCT were carried out at room with zero human movements. RSS values here indicated nominal variations (see Fig. 3) at normal room temperature of 20^{o}c. Table I lists the recorded observations during orientation and alignment experiments.

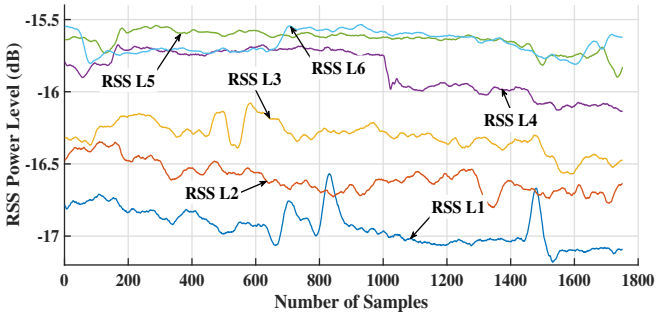


Fig. 3. Pilot RSSI Variations under Normal Conditions (Devoid of any Object in between TX and RX),and the "L" denotes Six observed levels of RSS. [32].

The NI-USRP based TX and RX orientation results revealed that TX and RX pair, with individual gains set to 50, show an optimal transmission characteristics at 0^{o}, with both the devices placed one meter (1m) above the ground level. The RSS drastically reduced to the minimum level at alternate orientations, and only maximized at 0 degrees. Therefore, LOS orientation of TX and RX established a direct communication link for our system model.

TABLE I. OBSERVED VALUES FOR PROPER ORIENTATION OF TX AND RX

TX-RX distance	TX Gain	RX Gain	Orientation (degrees)	RSS (dBm)
5 m	50	50	0	-15 dBm
			90	-29 dBm
			180	-36 dBm
			270	-34 dBm
5 m	50	40	0	-25 dBm
			90	-33 dBm
			180	-44 dBm
			270	-42 dBm
5 m	50	30	0	-31 dBm
			90	-42 dBm
			180	-52 dBm
			270	-51 dBm
5 m	50	20	0	-43 dBm
			90	-54 dBm
			180	-66 dBm
			270	-63 dBm
5 m	50	10	0	-53 dBm
			90	-65 dBm
			180	-75 dBm
			270	-73 dBm

TABLE II. OPTIMAL GAIN SETTING FOR TX AND RX

TX and RX distance	TX Gain	RX Gain	RSS (dB) (Noise)
5 m	0	0	-115 dB
5 m	0	50	-78 dB
5 m	50	0	-68 dB
5 m	50	50	-15 dB

3) Dataset distribution: Two individual datasets collected for each participant enabled creation of Training and Test datasets. Both the datasets contained 10872 data samples acquired during 90 seconds. Thus, for four participants, the train CSV file contained 43488 samples. However, sequentially placed moves in the test set were used for estimating the classification accuracy of the selected ML model. The raw datasets were transformed into Comma-separated values (CSV) files since the ML model in LabVIEW required CSV type input. All the data sets were class labelled manually, and each dataset contained six numbers corresponding to six different classes of moves. Next, human motion data in CSV files – train and test, were input to the SVM classifier of LabVIEW for evaluating the training and test accuracies respectively.

4) Data cleansing and anomaly correction: During the experiment, high frequency random noise was observed in all the four collected datasets. To filter out unwanted noise and glitches, a low pass filter and two consecutive moving average filters with a window size of 50 were applied on datasets. Effects of removing high frequency components and random noise can be seen in Fig. 4 and Fig. 5, with Fig. 4 showing a move set with added random noise, whereas a cleansed dataset can be observed in Fig. 5 after applied filtering. Abreast removing noise content, unusual movements such as hand gestures, or leaning backwards on a wall, or turning around for the next move, were observed, however, these inconsistencies were manually removed by overwriting the unwanted samples with mean variations.

For each move of each subject, RSS motion patterns embedded distortion and severe noise components. Each motion signature also showed anomalies resulting in trends in the data patterns. RSS strength drastically reduced with taller participants, leading to a change in scale of measurement. Thus, before RSS data was input to the SVM algorithm, processing was carried out at an earlier stage. Data processing

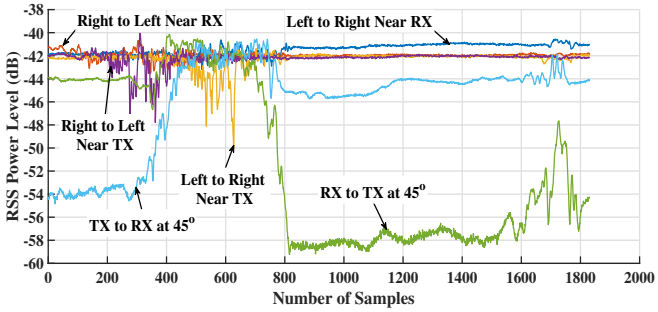


Fig. 4. Raw Movement Patterns before Filtering.

involved computing mean, standard deviations, detrending, normalization, and windowing to extract the required motion pattern only. Fig. 9 to 12, depict the extracted motion patterns. Notice some motion signatures contain unusual variations ranging over the last 300 samples. Variations as such represent participants movements such as turning around, raising arm, and leaning. These undesired data variations were manually removed, and finally cleaned motion signatures were acquired for human motion predictive analysis.

IV. HUMAN CLASSIFICATION USING MOTION SIGNATURES

The study classifies humans based on identified unique motion patterns observed in variations revealed by pilot signal. Recognizing and classifying human motion requires designating a class to each unique move dataset of each participant. Therefore, machine learning based motion recognition and classification algorithms are realized. Motion patterns observed from pilot signal variations exhibit nonlinear behavior, and Support Vector Machine (SVM) algorithms are mostly applicable to such applications, enabling exploration of hidden patterns in linear as well as non-linear data [33]. SVMs employ support vectors set out of training data to classify any unknown data sample q by comparing given input samples against the support vectors:

$$\text{sign}\left(\sum y_i \alpha_i K(p_i, q) + b\right) \tag{3}$$

where y_i represents class association (-1 or +1); α_i is the weight coefficient or Lagrange multiplier; K is the kernel function; p_i is the support vector data; i is the index from $i = 1, 2, 3, \ldots l$; and b represents the hyperplane distance from the origin. Next, subsections outline the variations of SVM algorithms.

A. SVM - Multiple Class input Categorization

Classification using the SVM algorithm typically requires defining a minimum of two classes or categories. Classes exceeding more than two, directs SVM algorithms to implement a one-versus-one approach for generating a binary classification model that corresponds to every possible class combination. Abreast, SVM algorithm implements polling method to derive most suitable class for a known input. For multiple classes resulting out of a polling method, the algorithm determines the class nearest to the sampled input. Thus, for the s number of classes generates $s \times (s-1)/2$ classification models, enabling a contrast in each category of input data.

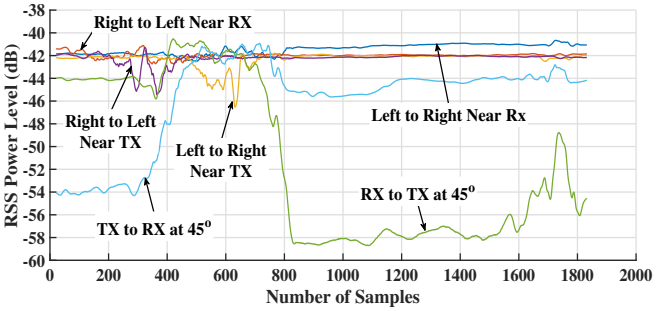


Fig. 5. Movement Patterns after Filtering.

B. Optimum Model Selection

SVM model determines classification of data samples, and classification problems involving single and multiple classes in the input data, employ either one-class model or multi-class models, known as C-SVC or nu-SVC.

1) SVM – Type C-SVC: To minimize the estimated error function, C-SVC model targets segregation of data samples separated by close or narrow margins using trained C-SVC model:

$$\min_{z,b,\xi} \quad \frac{1}{2} z^T z + \alpha \sum_{i=1}^{l} \xi_i \tag{4}$$

Subject to

$$y_i \left(z^T K\left(p_i\right) + b \right) \geq -\xi_i; \qquad \xi_i \geq 0, \quad i = 1, 2, \ldots l \tag{5}$$

where z, α and ξ represent normal vector of the hyperplane to origin, cost parameter, and the slack variable respectively. α – the Cost parameter uses partial training errors to define new soft margins subject to SVM algorithm failing to set out a an optimized margin. Selecting high values of *cost* parameter enables partial error removal, thereby resulting in a narrower margin and perfect classifications.

2) SVM – Type Nu-SVC: SVM class comparisons using *Nu*-SVC model enable precise controlling of training errors and support vectors set, using a parametric control called nu. Nu-SVC model requires training with input data to minimize the error function:

$$\min_{z,b,\xi} \quad \frac{1}{2} w^T w - v\lambda + \frac{1}{l} \sum_{i=1}^{l} \xi_i \tag{6}$$

Subject to

$$Y_i \left(z^T K\left(x_i\right) + b \right) \geq \rho - \xi_i; \ \xi_i \geq 0, \ i = 1, 2, \ldots l; \lambda \ \geq 0 \tag{7}$$

where v and ξ represent nu parameter and the slack variable, respectively. nu $(0 \leq nu \leq 1)$ parameter specifies maximum training errors ratio and minimum support vector count corresponding to sample count. Abreast increasing acceptance of texture defects, high value to nu increases probability of acceptance of texture dissimilarities.

The C-SVC classification model is used and a multi-featured vector data sets of four participants with associated label sets is prepared. During training, training data is manually labelled, however, testing determines the classes from the model itself.

C. Kernels

The SVM classifier comes with different kernel types. One of the kernels is categorized as a linear classifier, generally implements a linear kernel as a product of the input sample feature vector times the sample support vector, however, SVMs also support non-linear type of classifiers. Table III shows the most commonly used nonlinear kernels in SVM classifiers.

TABLE III. KERNEL TYPES [34]

Kernel type	Model Equation
Linear	$\text{Kernel}\left(x_i, x\right)$
Polynomial	$(\gamma \times \text{Kernel}\left(x_i, x\right) + \text{Coefficient})$
RBF	$e^{-\gamma \ \left\| x_i \ - \ x \right\|^2}$
Sigmoid	$e^{-\frac{\left\| x_i - \ x \right\|^2}{2 \times \sigma^2}}$

D. Feature Extraction

Feature extraction required preparing feature vectors out of the collected pilot signal variation datasets of all the four participants. The extracted feature vector space consisted of RSS variations for each participant observed in four directions, including features such as mean, variance, standard deviation, and skewness. For four participants, datasets $\left(p_i^1, q_j^1\right)$, $\left(p_i^2, q_j^2\right)$, $\left(p_i^3, q_j^3\right)$ *and* $\left(p_i^4, q_j^4\right)$ were collected, processed and then manually assigned class labels 1,2,3 and 4, corresponding the to four different participant, respectively (superscripts and superscripts denote index of participant and collected data points, respectively). The extracted feature vectors can be expressed in matrix as:

$$\begin{pmatrix} (p_{1,1}^1) & (p_{1,2}^1) & (p_{1,3}^1) & \cdots & (p_{1,n}^1) & \rightarrow (q_1^1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (p_{i,1}^1) & (p_{i,2}^1) & (p_{i,3}^1) & \cdots & (p_{i,n}^1) & \rightarrow (q_1^1) \\ \hline (p_{1,1}^2) & (p_{1,2}^2) & (p_{1,3}^2) & \cdots & (p_{1,n}^2) & \rightarrow (q_2^2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (p_{i,1}^2) & (p_{i,2}^2) & (p_{i,3}^2) & \cdots & (p_{i,n}^2) & \rightarrow (q_2^2) \\ \hline (p_{1,1}^3) & (p_{1,2}^3) & (p_{1,3}^3) & \cdots & (p_{1,n}^3) & \rightarrow (q_3^3) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (p_{i,1}^3) & (p_{i,2}^3) & (p_{i,3}^3) & \cdots & (p_{i,n}^3) & \rightarrow (q_3^3) \\ \hline (p_{1,1}^4) & (p_{1,2}^4) & (p_{1,3}^4) & \cdots & (p_{1,n}^4) & \rightarrow (q_4^4) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (p_{i,1}^4) & (p_{i,2}^4) & (p_{i,3}^4) & \cdots & (p_{i,n}^4) & \rightarrow (q_4^4) \end{pmatrix} \tag{8}$$

Where i represents a sample index taking value $i = 1 \rightarrow 1812$ for each subset matrix. $j = 1 \rightarrow 4$. $p_{1,1}^1$ to $p_{i,n}^1$ represent a subset matrix representing a dataset comprising pilot samples p^1 including assigned label vector q^1 for each participant. Similarly, three participants' feature vectors $p_{1,1}^2$ to $p_{i,n}^2$, $p_{1,1}^3$ to $p_{i,n}^3$ and $p_{1,1}^4$ to $p_{i,n}^4$ are designated labels q_2^2, q_3^3 and q_4^4, respectively.

E. Training – SVM Model

The SVM model was generated using the randomized optimization algorithm (ROA) available in LabVIEW. ROA running on the laptop equipped with Intel Core I3 processor and 8 GB of RAM, enabled fine tuning and optimization of SVM model parameters. After 250 iterations, the generated SVM model depicted approximately 99% training accuracy.

Observed training results on selected kernel types are given in Table IV. Both C-SVC and Nu-SVM models showed comparative results, however, SVM model of type C-SVC using Radial Bias Function (RBF) kernel resulted in nominal parameter settings with highest efficiency among all (see Table IV entry at row 3). With lowest prediction error against, the competitors (linear, polynomial, and sigmoid) guided us to select the C-SVC model with RBF kernel type, for this study. Results in Table IV explain that the RBF kernel function shows the lowest classification error. Fig. 6 to 9 depict each participants' raw motion pattern datasets, embedded with unwanted noise components, emerging as a result of pilot signal variations.

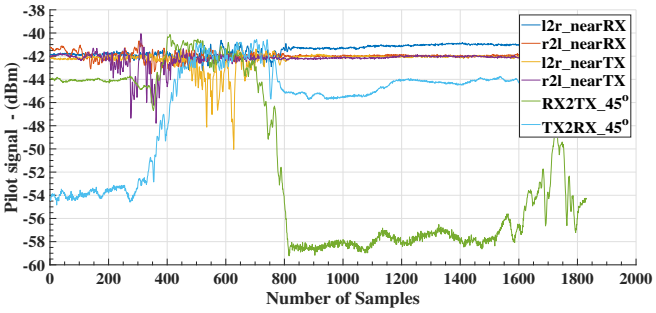


Fig. 6. Raw Motion Patterns of Participant – 1.

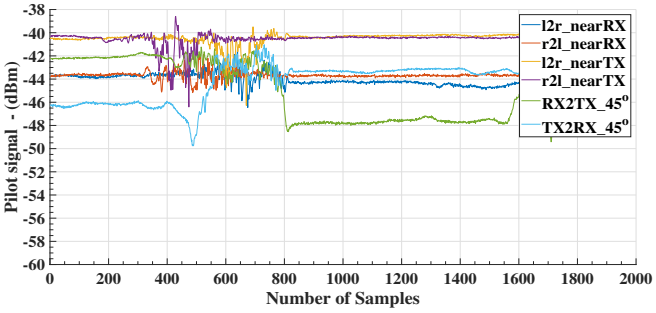


Fig. 7. Raw Movement Patterns of Participant – 2.

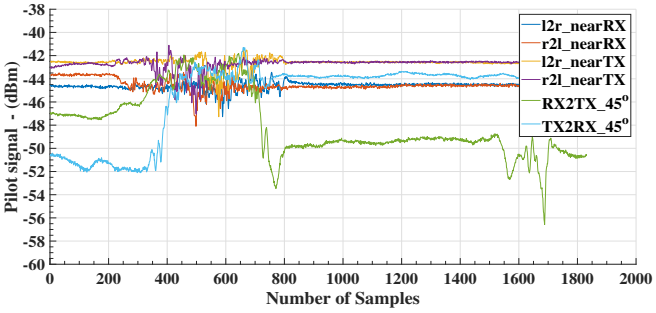


Fig. 8. Raw Movement Patterns of Participant – 3.

F. Prediction – Testing the SVM Model

Test datasets prepared initially, were used to test the prediction accuracy of trained SVM models. Test set contained sequentially arranged moves, however, without class labels this time, and using an overall test dataset containing $1812 \times 4 = 7248$ (in each feature vector) samples in total allowed performance measurements. The overall test data set

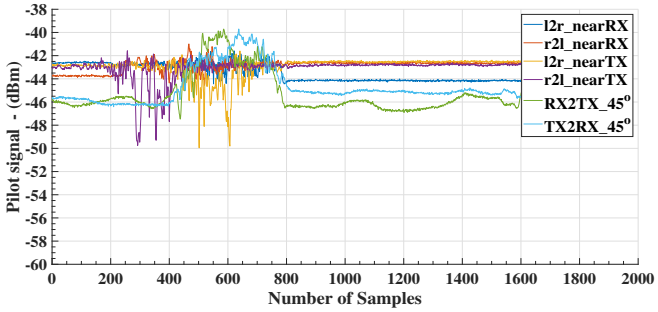


Fig. 9. Raw Movement Patterns of Participant – 4.

was used to measure the performance of the SVM model, which resulted in predicting each participant's identity with an accuracy of approximately 98%. Fig. 11 shows the screenshot of a human prediction application developed in LabVIEW, and the upper plot here shows four individual motion patterns input to the SVM algorithm, whereas lower plot shows the prediction results. The prediction result indicates some deviations in estimating accuracy of some samples. Abreast using the overall test dataset, each participant's motion dataset were isolated, and then tested using the trained SVM model individually on each participant's motion dataset. Here, the SVM model predicted each participant's identity with an accuracy of nearly 99%. Fig. 10 illustrates the True Positive Rate (TPR) and False Positive Rate (FPR) accuracies observed on the overall dataset containing moves of all the participants.

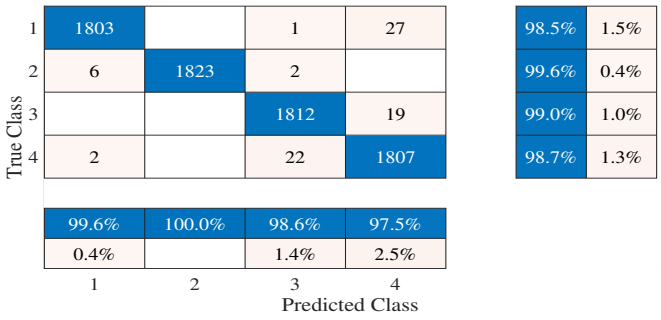


Fig. 10. Confusion Matrix of Overall Prediction Results.

V. Results and Discussion

Our testbed setup revealed numerous challenges, particularly with the lower strength of the pilot signal. The NI-USRP 2901 RX, showed that the pilot signal is highly sensitive to ambient variations, and hence considerable variations can be observed in the pilot RSS, having fine human movements, such as turning of head, raising arm or speaking. Our experimental setup ensured zero human motion, during the installation and initial testing. Fig. 3 shows the pilot signal strength variations, recorded under 27^o room temperature, with Tx and RX gains both set to 70 . The six of the pilot signal variations, show how pilot RSS varies at the nominal room temperature, with zero human presence.

The pilot signal variations range between -15.4 dB to -17.5 dB approximately. When the testbed was exposed to detect the human movements, the pilot RSS showed drastic variations, mostly with a power level below -45 dBm, thereby presenting unique motion patterns via pilot signal. Thus, the

TABLE IV. A RANDOMIZED SEARCH OPTIMIZATION CRITERIA EVALUATED IN LABVIEW

Sno.	kernel	svm_type	c	nu	deg	gamma	coeff	accu
1	RBF	CSV-C	0.7	0	3	0.7	0	0.999
2	RBF	CSV-C	0.9	0	3	0.4	0	0.999
3	RBF	CSV-C	0.8	0	3	0.7	0	0.999
4	RBF	CSV-C	1	0	3	0.6	0	0.999
5	RBF	CSV-C	1	0	3	0.7	0	0.999
6	Polynomial	CSV-C	0.4	0	3	0.4	1	0.998
7	Polynomial	CSV-C	0.9	0	8	0.1	1	0.998
8	Polynomial	CSV-C	0.8	0	4	0.1	1	0.998
9	Polynomial	CSV-C	1	0	3	0.7	1	0.998
10	Polynomial	CSV-C	0.7	0	2	0.7	1	0.998
11	Polynomial	CSV-C	0.5	0	4	0.5	1	0.998
12	Polynomial	CSV-C	1	0	2	0.6	1	0.998
13	Polynomial	CSV-C	0.7	0	9	0.1	1	0.998
14	Polynomial	CSV-C	0.2	0	3	0.8	1	0.998
15	Polynomial	CSV-C	0.9	0	2	0.6	1	0.998
16	RBF	CSV-C	0.4	0	3	0.7	0	0.998
17	Polynomial	CSV-C	0.2	0	4	0.7	1	0.998
18	Polynomial	CSV-C	0.7	0	2	0.5	1	0.998
19	Polynomial	CSV-C	0.6	0	6	0.2	1	0.998
20	Polynomial	CSV-C	0.1	0	5	0.4	1	0.998
21	Polynomial	CSV-C	0.7	0	3	0.8	0	0.998
22	Polynomial	CSV-C	0.1	0	3	0.7	1	0.998
23	Polynomial	CSV-C	0.4	0	4	0.6	1	0.998
24	Polynomial	CSV-C	0.1	0	2	0.4	1	0.998
25	Polynomial	CSV-C	0.4	0	7	0.1	1	0.998
26	Polynomial	CSV-C	1	0	3	0.8	0	0.998
27	Polynomial	CSV-C	0.8	0	4	0.6	1	0.998
28	Polynomial	CSV-C	0.1	0	2	0.6	1	0.998
29	Polynomial	CSV-C	0.6	0	6	0.1	1	0.998
30	Polynomial	CSV-C	0.8	0	3	0.6	1	0.997
31	Polynomial	CSV-C	0.6	0	3	0.7	0	0.997
32	Polynomial	CSV-C	0.6	0	2	0.8	1	0.997
33	Polynomial	CSV-C	0.4	0	3	0.4	0	0.997
34	Polynomial	CSV-C	0.9	0	2	0.8	1	0.997
35	Polynomial	CSV-C	0.1	0	3	0.6	0	0.997
36	Polynomial	CSV-C	0.8	0	5	0.2	1	0.997
37	RBF	CSV-C	0.2	0	3	0.4	0	0.997
38	Polynomial	CSV-C	0.2	0	3	0.7	0	0.997
39	Polynomial	CSV-C	0.9	0	4	0.5	1	0.997
40	Polynomial	CSV-C	0.9	0	4	0.7	1	0.997
41	Polynomial	CSV-C	1	0	4	0.7	1	0.996
42	Polynomial	CSV-C	0.5	0	2	0.1	1	0.996
43	Polynomial	CSV-C	1	0	4	0.5	1	0.996
44	Polynomial	CSV-C	0.6	0	2	0.1	1	0.996
45	Polynomial	CSV-C	0.1	0	3	0.3	0	0.996
46	Polynomial	CSV-C	1	0	6	0.3	1	0.996
47	Polynomial	CSV-C	0.5	0	5	0.8	1	0.996
48	Polynomial	CSV-C	0.5	0	4	0.7	0	0.996
49	Polynomial	CSV-C	0.5	0	5	0.7	1	0.996
50	Polynomial	CSV-C	0.9	0	3	0.2	0	0.996

proposed human motion recognition and identification method is highly sensitive, and our analysis of collected results (see Fig. 12 to 17) suggests that fluctuations in pilot RSS due to human motion can be extracted and analyzed to uncover unique human motion patterns for recognizing a person's identity. Therefore, our study results are applicable to environments such as smart authentication systems, or patient monitoring systems, and healthcare monitoring systems, wherein passive sensing is mostly preferred over traditional methods, typically based on active sensing devices.

1) Comparative Analysis: SVM algorithm prediction accuracy was estimated on four different kernel functions – linear, radial basis function (RBF), sigmoid and polynomial. Using grid-search optimization, setting different input parameters, as required by the kernel functions, enabled the prediction model to achieve maximum classification accuracy level [35]. Table IV defines the first 50 configuration parameters used for tuning the SVM prediction model. As shown in the Table IV, the SVM type C-SVC models using a polynomial kernel reveals the most efficient hyperplane model. The CSV-C kernel with RBF kernel implemented in labVIEW showed 99.9% training accuracy, whereas classification accuracy of human identification (participant identity prediction) on the train dataset was observed to be approximately 98% with polynomial kernel. CSV-C with sigmoid type kernel is strongly competitive with the other CSV-C types, however, resulting in a prediction accuracy of just below 95%. Fig. 18 and 19 show the probability distribution of training accuracy levels realized against the different kernels and SVM model types respectively. The CSV-C type RBF kernel is observed to deliver higher prediction accuracy than other kernel types. CSV-C with linear kernel was realized to be least efficient as compared to the sigmoid kernel, which showed a mean accuracy level at 91%.

Estimation accuracy of linear kernel shows the least probability distribution, while the sigmoid shows 83%, 87%, 91%, 94% and 97%, for minimum, lower quartile, median, upper quartile and maximum distributions respectively. Among all the four, the polynomial kernel spans lower distribution, and shows lower and upper quartiles at 95% and 98% respectively, affirming that polynomial has considerable training accuracy compared to the RBF kernel with overlapped upper and lower

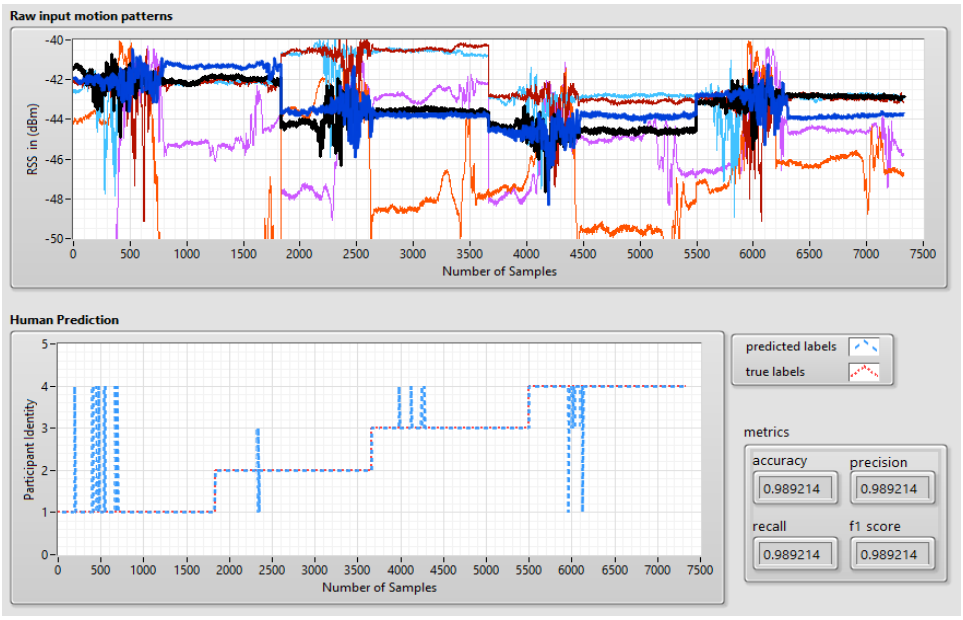


Fig. 11. LabVIEW: SVM Prediction Results on Test Dataset of all Participant.

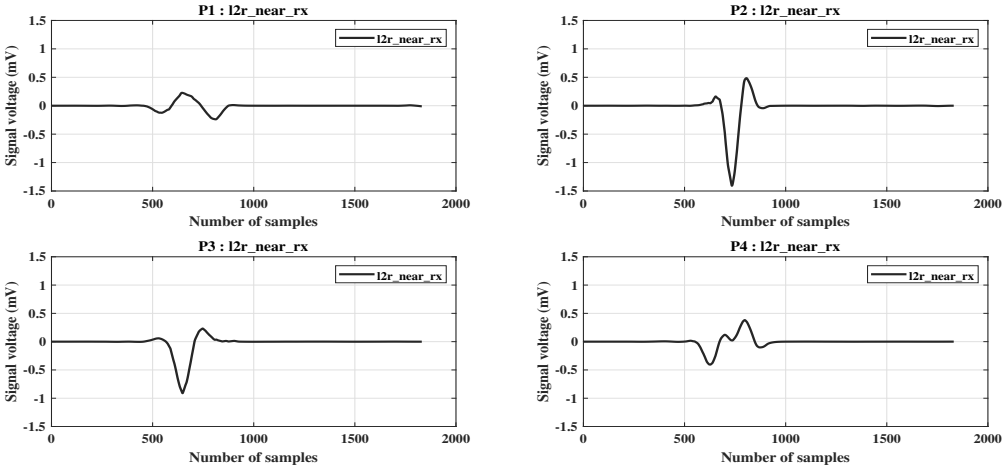


Fig. 12. Processed Move(1) Motion Signatures.

quartiles. On the other hand, sigmoid kernel has a more wider distribution, spanning approximately from 88% to 95%.

Observing the boxplot of SVM model types, describes higher efficiency of CSV-C on non-linear human motion dataset with lower and upper quartiles, both above the 95% accuracy level. Although Nu-SVC showed wider accuracy distribution than its counterpart CSV-C, the observed accuracy range on the non-linear human motion dataset is still lower than CSV-C, with both lower and upper quartiles below the mean of CSV-C type kernel. This clearly shows the CSV-C is suitable for applications involving non-linear data distributions, in particular human motion recognition using highly sensitive pilot signal RSS.

2) Testbed requirements: NI-USRP devices may use different hardware clock signal generators. Often this raises critical concerns since mismatch in clock signal frequency can result in the receiver showing incorrect data. The NI-USRP devices employed by this study, showed deviating results due to mismatch in their clock signals. Since the devised system needed recognizing motion patterns in microsecond time, the USRPs experienced jitter including pilot signal frequency drift to $\pm$ 4 KHz. Thus, pilot signal appeared 10 KHz $\pm$ 4 KHz, and to abase frequency offsets in the pilot signal, manually changing sampling frequency in NI-USRP devices enabled tuning the receiver (RX) to show correct pilot signal frequency response, which in our case was 10 KHz. Thus, tuning of both the TX/RX parameters, such as transmitted power, sampling frequency, pilot signal frequency, distance and receiver orientation, ensured optimal system behaviour. Although sampling rate is adjusted manually in this study, use of an external reference clock generator is recommended for future studies.

Moreover, this study employed NI-USRP 2901, connected

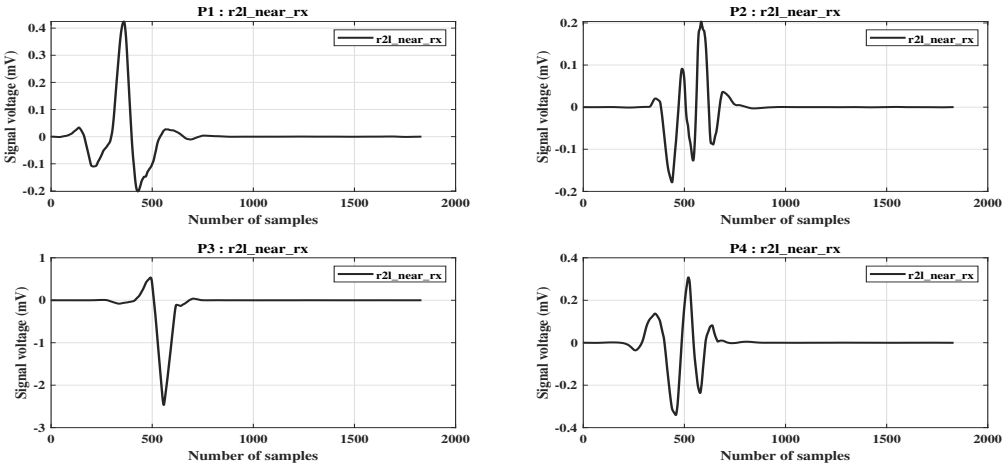


Fig. 13. Processed Move(2) Motion Signatures.

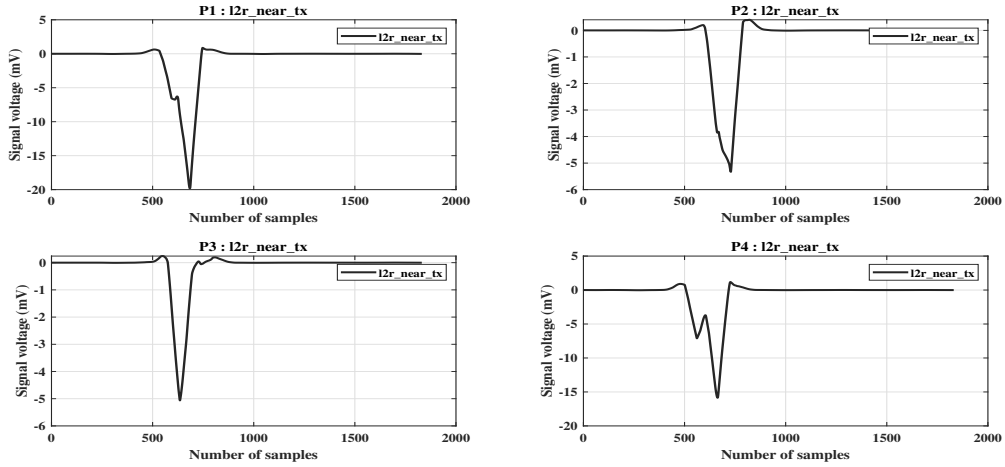


Fig. 14. Processed Move(3) Motion Signatures.

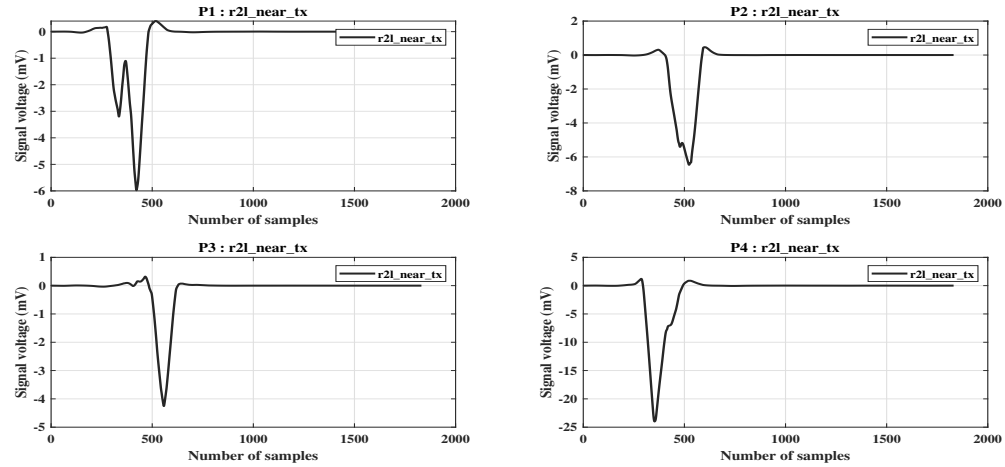


Fig. 15. Processed Move(4) Motion Signatures.

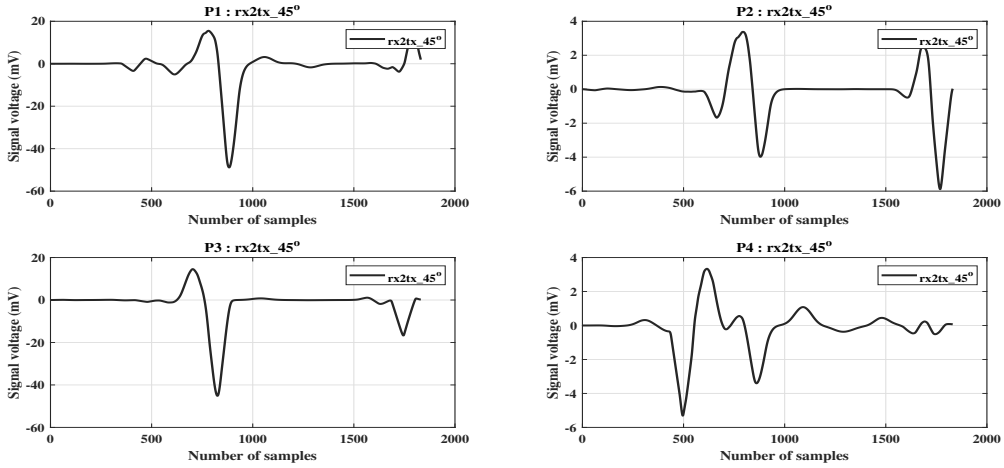


Fig. 16. Processed Move(5) Motion Signatures.

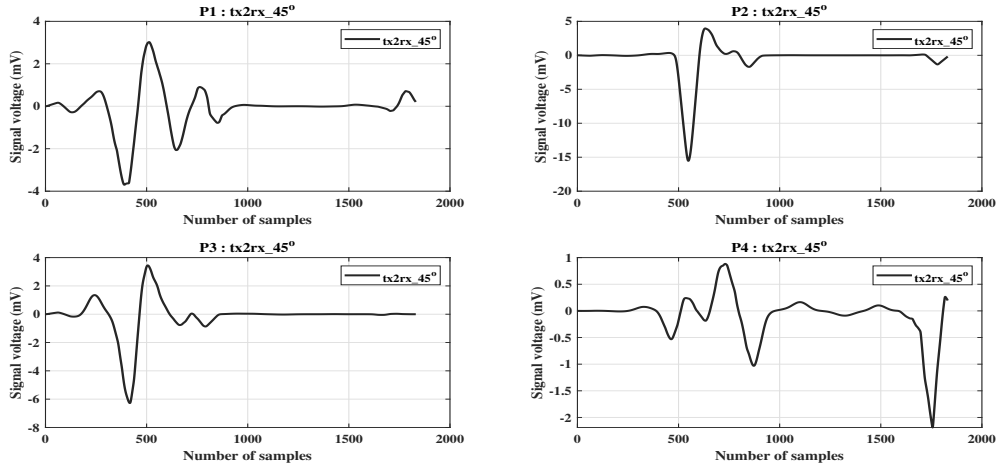


Fig. 17. Processed Move(6) Motion Signatures.

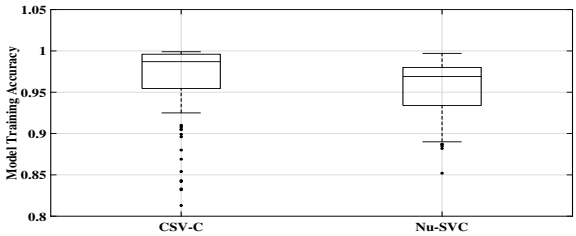


Fig. 18. Boxplot of Training Accuracy versus Kernel Types.

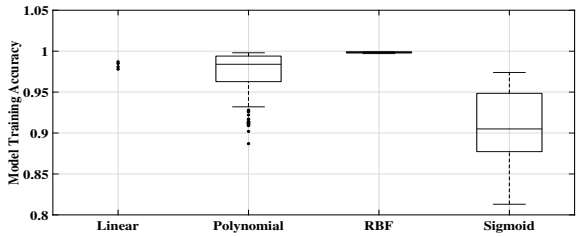


Fig. 19. Boxplot of Training Accuracy versus SVM Types.

to a host computer via USB-3.0 port. Despite the high speed access offered by the USB-3.0 communication protocol and wideband capabilities (from 70 MHz to 6 GHz) of NI-USRP 2901, the LabVIEW communication design software application posed start up delays, and even during the device initialization phase, often lag to communicate with the host device, the NI-USRP. Therefore, for future research work, high speed devices such as NI-USRP 2921 [30], offering 1 Gigabit Ethernet speeds is highly recommended since real time processing requires instant processing, and communication devices must ensure higher data throughput. With the manually tuned testbed, our observed results indicate high precision in the training and test accuracies compared to the accuracies observed with multi sub-carrier based human motion detection systems.

VI. Conclusion

This study proposed a human motion recognition and identification system, implementing a single sub-carrier CW and SVM based pattern recognition system. The study employed a testbed for human motion data acquisition using state-of-the-art Software Defined Radios – NI-USRP 2901 and LabVIEW software. Operating at 5GHz frequency, a SISO channel created by transmitter (TX) and receiver (RX) enabled human motion patterns analysis using Support Vector Machine algorithm (SVM)to uncover human identities. Experiment was carried out in a controlled environment enabling us to set up environmental parameters as well as to fine tune the custom designed transmitter receiver. The findings in this study reveal that highly sensitive pilot RSS and SVM algorithm information can significantly help in recognizing human motion, which in turn guides to human identification (4 subjects) with the prediction accuracy of 98.92%. Our proposed testbed system can be devised and validated on any of the commercially available NI-USRP added with LabVIEW Communication Design Software Suite.

The study anticipates a live motion recognition system in the future works, prototyping the current setup on an embedded device. SDRs in association with the high speed FPGA, support running SVM models directly. Therefore, this will serve as a guide to researchers interested in developing real time human motion recognition testbeds, thereby enabling them to test and evaluate system performances in other available ISM bands. Envisioned here is the development of more efficient, robust, and highly accurate human motion recognition systems, while using pilot signal RSS can also in human motion direction identification, still remains to be an open research problem.

Acknowledgment

This work was partially supported by the Kuwait Foundation for Advancement of Sciences (KFAS) under Grant #PR-15NH-04. The paper adds a part of our previously published paper in the – 2020 IEEE Wireless and Networking Conference (WCNC).

References

[1] N. Patwari and J. Wilson, "Rf sensor networks for device-free localization: Measurements, models, and algorithms," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1961–1973, 2010.

[2] K. Witrisal, P. Meissner, E. Leitinger, Y. Shen, C. Gustafson, F. Tufvesson, K. Haneda, D. Dardari, A. F. Molisch, A. Conti, and M. Z. Win, "High-accuracy localization for assisted living: 5g systems will turn multipath channels from foe to friend," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 59–70, 2016.

[3] M. Chen, K. Liu, J. Ma, X. Zeng, Z. Dong, G. Tong, and C. Liu, "Moloc: Unsupervised fingerprint roaming for device-free indoor localization in a mobile ship environment," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[4] ——, "Moloc: Unsupervised fingerprint roaming for device-free indoor localization in a mobile ship environment," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[5] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, 2012.

[6] F. Fereidoonian, F. Firouzi, and B. Farahani, "Human activity recognition: From sensors to applications," in *2020 International Conference on Omni-layer Intelligent Systems (COINS)*, 2020, pp. 1–8.

[7] X. Zhou, W. Liang, K. I. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for internet of healthcare things," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6429–6438, 2020.

[8] K. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.

[9] M. Bennamoun, Y. Guo, F. Tombari, K. Youcef-Toumi, and K. Nishino, "Guest editors' introduction to the special issue on rgb-d vision: Methods and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2329–2332, 2020.

[10] R. Mondal, D. Mukherjee, P. K. Singh, V. Bhateja, and R. Sarkar, "A new framework for smartphone sensor based human activity recognition using graph neural network," *IEEE Sensors Journal*, pp. 1–1, 2020.

[11] J. Lu, X. Zheng, M. Sheng, J. Jin, and S. Yu, "Efficient human activity recognition using a single wearable sensor," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[12] O. Barut, L. Zhou, and Y. Luo, "Multi-task lstm model for human activity recognition and intensity estimation using wearable sensor data," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[13] M. Zhou, Y. Wang, Z. Tian, Y. Lian, Y. Wang, and B. Wang, "Calibrated data simplification for energy-efficient location sensing in internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6125–6133, 2019.

[14] K. Woyach, D. Puccinelli, and M. Haenggi, "Sensorless sensing in wireless networks: Implementation and measurements," in *2006 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, 2006, pp. 1–8.

[15] M. Youssef, M. Mah, and A. Agrawala, "Challenges: Device-free passive localization for wireless environments," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, ser. MobiCom '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 222–229. [Online]. Available: https://doi.org/10.1145/1287853.1287880

[16] T. Chang, L. Wang, and F. Chang, "A solution to the ill-conditioned gps positioning problem in an urban environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 135–145, 2009.

[17] B. Yang, L. Guo, R. Guo, M. Zhao, and T. Zhao, "A novel trilateration algorithm for rssi-based indoor localization," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 8164–8172, 2020.

[18] P. Zhang, Z. Su, Z. Dong, and K. Pahlavan, "Complex motion detection based on channel state information and lstm-rnn," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 0756–0760.

[19] Z. Zhang, W. Nie, Y. Wang, and L. Xie, "Channel state information based indoor localization error bound leveraging pedestrian random motion," *IEEE Access*, vol. 8, pp. 153 311–153 321, 2020.

[20] M. T. Islam and S. Nirjon, "Wi-fringe: Leveraging text semantics in wifi csi-based device-free named gesture recognition," in *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2020, pp. 35–42.

[21] Z. Han, L. Guo, Z. Lu, X. Wen, and W. Zheng, "Deep adaptation networks based gesture recognition using commodity wifi," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–7.

[22] Y. Chen, R. Ou, Z. Li, and K. Wu, "Wiface: Facial expression recognition using wi-fi signals," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020.

[23] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, 2017.

[24] J. Yang, H. Zou, H. Jiang, and L. Xie, "Carefi: Sedentary behavior monitoring system via commodity wifi infrastructures," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7620–7629, 2018.

[25] Z. Fu, J. Xu, Z. Zhu, A. X. Liu, and X. Sun, "Writing in the air with wifi signals for virtual reality devices," *IEEE Transactions on Mobile Computing*, vol. 18, no. 2, pp. 473–484, 2019.

[26] H. Abdelnasser, K. A. Harras, and M. Youssef, "A ubiquitous wifi-based fine-grained gesture recognition system," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2018.

[27] L. Zhang, Q. Gao, X. Ma, J. Wang, T. Yang, and H. Wang, "Defi: Robust training-free device-free wireless localization with wifi," *IEEE Trans. on Vehicular Tech.*, vol. 67, no. 9, pp. 8822–8831, 2018.

[28] S. D. Domenico, M. D. Sanctis, E. Cianca, F. Giuliano, and G. Bianchi, "Exploring training options for rf sensing using csi," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 116–123, 2018.

[29] O. A. http://www.ni.com/en-lb/support/model.usrp 2901.html, "Ni usrp 2901."

[30] J. Mitsugi, Y. Kawakita, K. Egawa, and H. Ichikawa, "Perfectly synchronized streaming from multiple digitally modulated backscatter sensor tags," *IEEE Journal of Radio Frequency Identification*, vol. 3, no. 3, pp. 149–156, 2019.

[31] K. Wu, Jiang Xiao, Youwen Yi, Min Gao, and L. M. Ni, "Fila: Fine-grained indoor localization," in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 2210–2218.

[32] S. A. Bhat, A. Mehbodniya, A. E. Alwakeel, J. Webber, and K. Al-Begain, "Human motion patterns recognition based on rss and support vector machines," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.

[33] W. Buwei, C. Jianfeng, W. Bo, and F. Shuanglei, "A solar power prediction using support vector machines based on multi-source data fusion," in *2018 International Conference on Power System Technology (POWERCON)*. IEEE, 2018, pp. 4573–4577.

[34] N.-E. Ayat, M. Cheriet, and C. Y. Suen, "Empirical error based optimization of svm kernels: Application to digit image recognition," in *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*. IEEE, 2002, pp. 292–297.

[35] Y.-D. Zhang and L. Wu, "Classification of fruits using computer vision and a multiclass support vector machine," *Sensors (Basel, Switzerland)*, vol. 12, pp. 12 489–505, 12 2012.

Agile Fitness of Software Companies in Bangladesh: An Empirical Investigation

M M Mahbubul Syeed[1], Razib Hayat Khan[2]
Professional Member
ACM

Jonayet Miah[3]
North-South University
Bangladesh

***Abstract*—With the mandate of light-weight working practices, iterative development, customer collaboration and incremental delivery of business values, Agile software development methods have become the de-facto standard for commercial software development, worldwide. Consequently, this research aims to empirically investigate the preparedness and the adoption of agile practices in the prominent software companies in Bangladesh. To achieve this goal, an extensive survey with 16 established software companies in Bangladesh is carried out. Results exhibit that the Scrum agile methodology is the highest practiced one. Alongside, to a great extent these software companies have the readiness to effectively adopt the Scrum methodology. However, with regard to practicing the Scrum principles, they fall short in many key aspects.**

***Keywords*—Agile manifesto; agile methodology; scrum; software development projects; software companies in Bangladesh**

I. Introduction

For the past two decades, agile software development methods have becoming the de-facto standard worldwide for developing cutting age software systems [1]. Several variations of this method, e.g., Scrum, Extreme Programming, Crystal, FDD and others have attracted a lot of attention to the software engineering and research communities.

A group of agile practitioners, loosely known as *Agile Alliance* formulates the agile principles in 2001 [1]. These principles, popularly termed as *Agile Manifesto*, help to optimize the software development process and increase efficiency with greater customer satisfaction [2]. The Agile Manifesto provides the four core values for software development projects [1][2], namely, (a) Individuals and interactions over processes and tools, (b) Working software over comprehensive documentation, (c) Customer collaboration over contract negotiation, and (d) Responding to change over following a plan. Therefore, agile principles shifts the software development paradigm from plan-driven to value-driven process models [1][3].

Based on the agile manifestation, all agile methodologies at their very core implements rapid and iterative development process for continuous and incremental software delivery, have flexibility to accommodate changing requirements and market demands, and integrate customer feedback [5][7].

Statistics on the adoption and usage of Agile methods to run software development projects has shown overwhelming acceptance worldwide [18] [19]. It has been reported that around 70% of the companies, practice agile methods for software development[18]. According to 1015 developers around the world, agile practices are the integral part of their everyday activities [18]. Other studies, reported that agile projects are 28% more successful than traditional projects [19]. Alongside, in the USA, the average salary of an agile project manager is more than USD 90K [19].

Bangladesh being an emerging economy is rapidly extending Her presence in the world software market with a current market value of USD 130 billion [20]. In recent decade, several of Her software companies has accomplished a number of outsourced projects and thereby gaining reputation [20] [21]. With reference to this, the software companies in Bangladesh must demonstrate the authentic adoption and practice of Agile principles, norms, and practices to persuade their international clients and extend their market share even further.

This research aims to find out the extent to which the prominent software development companies in Bangladesh follow the Agile principles. Consequently, the primary contributions of this research is as follows: (a) empirically investigate into the Agile development practices in the context of established software companies located in Bangladesh, (b) analyze and comprehend the fitness of these software companies in relation to Agile practices, and (c) offer guidelines / scope of improvements based on the standards defined by Agile manifesto. Alongside, this reporting also assist the overseas cooperates to decide on outsourcing projects in Bangladesh.

This paper is organized as follows, in Section II the background work and focus of this study is presented, Section III detail the realization of the survey method for conducting this study. Result and recommendations are presented in Section IV. Finally, overall assessment, future works and concluding remarks are drawn in Sections V and VI, respectively.

II. Background and Research Focus

Agile software development principles are initially proposed and promoted by a group of 17 software professionals popularly known as the *Agile Alliance* [3]. They stated principles, norms, and practices for a set of *lightweight* software development methods in the form of *Agile Manifesto* [4] [3]. Thereafter, several Agile methods have been matured and put into practice. Among then, Scrum [5], Extreme Programming (XP) [6], Feature Driven Development (FDD) [12], Crystal [6], Lean Software Development [7], and Kanban [8], are the most common methods in the software industries [9].

Agile methods follow light-weight working practices, continuous development and delivery, integration of changing requirements and customer collaboration throughout the development process, over long-planning, cumbersome documenta-

tion, and inflexible development phases [10]. Therefore, these methods ensure high customer satisfaction through the delivery of business values in short iterations and incrementally with the option of accommodating changing needs even late within the development process [11] [4] [3].

Since their emergence, agile methods are used by more than 70% companies in their software development projects [18] [19]. Therefore, research related to the adoption and practice of agile methodologies in software companies has been the center concern in software engineering research. In [16], a survey based comparative study was conducted to find out the most popular agile methodologies practiced in the industries. Result suggests higher popularity of Scrum than that of Extreme Programming and Kanban. Alongside, the applicability and implication of agile development methods were investigated in [17].

In [13], an approach to effectively adopt agile methods, specially, Scrum is presented. A survey based research was conducted in [14] to formulate the challenges for enterprises to adopt agile methods. Reported results highlighted that there is no single agile method that can be universally applied, and have to be tailored to integrate into existing processes. On the track, a framework termed *Agile Software Solution Framework* were proposed and empirically verified to assist the companies in defining and introducing agility in the development process [15].

However, to the best of our knowledge, no comprehensive investigation has been reported to verify the extent to which software companies adopt and practice the agile principles, specially concerning the developing countries, e.g., Bangladesh.

Therefore, the primary focus of this research is three fold: (a) empirically investigate into the Agile development practices in the context of established software companies located in Bangladesh, (b) analyze and comprehend the fitness of these software companies in relation to Agile practices, and (c) offer guidelines / scope of improvements for these companies based on the standards defined by Agile manifesto. Alongside this reporting also support the overseas cooperates to decide on outsourcing projects to Bangladesh.

III. Research Approach

To conduct this research, an extensive *survey* is carried out with the established software companies based in Dhaka city, the capital of Bangladesh. The *Survey Research Method* is the best suited for a research of this nature, because, it is a comprehensive method for collecting information to describe, compare or explain knowledge, attitudes, and behavior on a given domain [22][23].

The target audience of our survey is the software professionals of different ranks who are currently employed in various prominent software companies in Bangladesh. A total of 38 professionals participated in the survey from 16 different companies, a taxonomy of which is discussed in Section IV-A.

1) Survey construction: To construct the survey, four agile methods that are most practised in software firms in general are selected [9]. These methods are, Scrum [5], XP [6], FDD [12] and Crystal [6]. Thereafter, based on the mandate and practices of each of the methods, specific set of questionnaires are designed. These questions further grouped into focused domains to better comprehend on the actual realisation of the methods within the companies. The questionnaire for Scrum method is detailed in Section VI (Fig. 14). For the other methods the questionnaires are omitted as they are not that popular according to our survey findings (discussed in Section IV-B). Alongside, to get the company and employee profiling, a common set of questions are also designed.

The questions have both close ended and open ended options to respond. The close ended options are developed in frequency scales, rather than two-point Yes/No scale. The usage of frequency scale has enabled to measure how frequently an event occurs when following a specific agile method. Furthermore, it helps to conduct statistical analysis from the data. The answer options for which frequency scale is used contained four options, they are: *none, rarely, sometimes, all the time*. The optional open ended part, allows the interviewee to complement their answer through the narrative expression.

The questions are kept short, to the point and unambiguous. Each question focuses on one aspect of the Agile method only. In formulating the questions, standard terms specific to each of the methods are used for greater clarity and understanding. Additionally, each question is associated with legends to further explain the content of the questions.

2) Survey execution : To execute this survey, an interactive Google form is designed with the questionnaire. This form is accompanied with the detail guidelines to assist the interviewees and navigate through the questionnaire session. The form begins with a common section to record company profile followed by four specific selections for the four selected methods. Based on the interviewee selection of the method, the corresponding method related questionnaire section is opened. Response is recorded in Google sheet categorically which is then extracted and analysed. To complete the survey, approximately 10 to 15 minutes of dedicated time is required.

The design of this survey is cross-sectional and are aimed at a fixed point of time. All the companies are contacted well before conducting the survey through official channel, and a *Non Disclosure Agreement (NDA)* was signed to maintain the secrecy and anonymity of the company specific information. The NDA also guarantee to some extent the accuracy of responses as the interviewee feels confident of not getting disclosed. Then the contact information (official emails and phone numbers) of the interviewees from each of the companies are collected. The survey form is sent over the email with clear guidelines and a follow up phone conversation is carried out in case of any clarification is required by the participant.

3) Evaluation approach: The survey instrumentation as prescribed above supports both quantitative and qualitative analysis on the collected data. The first part of the answer (frequency scale answers) allowed to get a generic perspective on a given aspect (e.g., *Do all the team members work in the same space?*) through quantitative investigation. To achieve this, related data are aggregated, grouped and charts are generated. The second part of the answer (i.e., the optional open ended response) is analyzed, comprehended and mapped with the corresponding charts to draw critical reasoning on the overall response. Once done, this assessment is verified against the

Firm	Year Est.	Age (Y)	Company Focus	Certification
Firm A	2001	19	**Product:** HR systems, Banking solutions. **Service:** Customer centric services	ISO 9001
Firm B	1998	21	**Product:** HR systems, Banking & Financial solutions, E-commerce, E-learning systems, Mobile applications **Service:** Testing as a service	CMMI L- 5 ISO 9001:2008
Firm C	2001	19	**Product:** IoT, Big Data, Deep Learning, Financial, Blockchain. **Service:** Offshore services, Application management as a service, Testing and automation	CMMI level 3 ISO 9001
Firm D	2006	14	**Product:** Banking solutions, Machine learning AI and cloud solutions, E-commerce sites, Game development, AR and VR systems	ISO 27001: 2013 9001: 2015
Firm E	2004	16	**Product:** Mobile Application (IOS and Android), Management software, Game development **Service:** Export software products overseas	ISO 9001:2008
Firm F	2001	19	**Product:** Healthcare systems, Banking solution, Management systems.	
Firm G	2003	17	**Product:** Mobile and Embedded applications for sectors like, entertainment, banking, insurance, pharmaceutical, telecommunication	
Firm H	2000	20	**Product:** E-commerce sites, Banking solutions **Service:** Largest internet service provider	ISO 9001:2008
Firm I	2010	10	**Product:** Mobile applications, E-commerce sites, Management systems	
Firm J	2003	17	**Product:** Mobile applications, Enterprise telecommunication solutions, E-governance, NLP, Machine learning **Service:** Operates in Singapore, Bangladesh, India, UK, USA, and Hong Kong	
Firm K	2010	10	**Product:** Mobile applications, E-commerce sites, Desktop software, Digital marketing.	
Firm L	2015	5	**Product:** Ridesharing platform, Food, Parcel, Courier services. **Service:** Fastest-growing tech startups in Asia.	
Firm M	2006	14	**Product:** Content management, Web applications, E-commerce sites, Game development, Project management and Accounting software. **Service:** Network monitoring and Administration	
Firm N	2009	11	**Product:** Mobile applications, Full-stack web development, Enterprise database design, UX/UI design **Service:** Providing full-stack resources to clients worldwide.	
Firm O	2017	3	**Product:** Web development, Mobile Application, Digital Marketing	
Firm P	2012	8	**Product:** Business software, ERP, HR management	

Fig. 1. The brief Portfolio of the Companies.

standard practices of the methods to derive recommendations.

4) Survey reliability: To ensure the reliability of the survey instrument, the so called *test-retest* approach is used. That is, the same respondents are surveyed once again at different point of time to observe the variation on the response.

IV. Analysis and Synthesis

This survey is conducted among 16 prominent software companies located in Dhaka city, the capital of Bangladesh. A total of 38 techno-professionals currently employed in these companies at different ranks have participated in this survey. In the following sections the transcript of evaluation is presented.

A. On the Company Profiling

The first part of the survey questionnaire is designed to get an overall portfolio of the software companies, especially focusing on their project focus, achieved standardisation, technology expertise, employee and project profiling. This taxonomy of company portfolio is required to assess the overall preparedness of the companies to carry out agile development while maintain all the key parameters to meet standard and quality [3] [9].

The brief portfolio of the 16 software companies is presented in Fig. 1. These companies have an average operational experience of 13.5 years with a maximum of 20 years and a minimum of 3 years (Column 2 and 3 of Fig. 1). There are 6 companies who already achieved ISO certification with two of them attaining Capability Maturity (CMM) level of 3 and 5 (Column 5 in Fig. 1). The average operational experience of these 6 companies is 18.2 years, therefore having a long trail of successful software project accomplishment. Their client base includes both local and international corporate and enterprises.

During their service life, most of these companies developed their expertise on both product development as well as service delivery on diverse categories, a classification of which

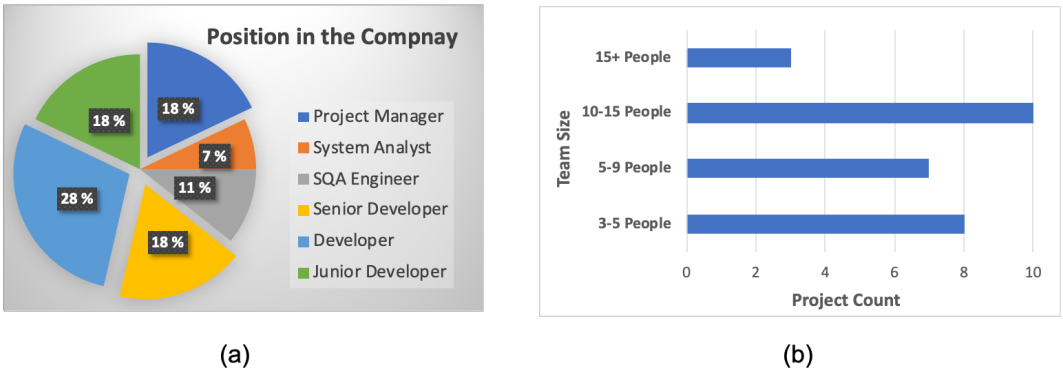


Fig. 2. (a) Employee Ranks and (b) Project Team Size within the Companies.

is presented in Fig. 3(a). According to this figure, E-commerce and Web services are the core focus followed by Management and Banking solutions. This observation is in line with the market demand [24]. Alongside, the mobile application and game development are cited as a major emerging market segment for these companies.

Accomplished projects have a development lifetime of either 3 to 6 months (very short to short duration), or 1/2 to 1 year (medium duration) or more than a year (long duration), depending on the requirements. A distribution of the projects along this lifespan reveals that 75% of the projects belong to medium and long duration with only 25% are from short duration. Additionally, the data support that the short duration projects are mostly performed by the new companies in the list having less than 10 years of experience in the field. Fig. 3(b) summarizes this observation.

All the companies offering a number of ranks to their employees that are typical for an established software development company to carry out their projects. According to the survey, 6 such ranks are offered, namely, *Project Manager, System Analyst, SQA Engineer, Senior Developer, Developer, and Junior Developer*. Fig. 2(a) narrates these ranks in a descending order with the proportion of each rank within the companies (calculated based on the total number of employees per rank in the 16 companies). As a reference to the reader, the *Project Manager* is the administrative lead for project planning, monitoring and managing the progress and resources. The *Systems Analyst* is the IT guru who is responsible to analyze the problem domain and to come up with the best approach in solving it. The *Senior Developer* is the highly experienced professional who lead a team of developers in getting the development work done. The *Developer* is responsible for messing up their hands with implementing the code by following best practiced design patterns. Part of their responsibility includes training and assigning development tasks to *Junior Developers* and assist them. Finally, the *SQA Engineer* is responsible for designing and executing the test plan and assist the development team to resolve them [25][26].

According to the statistics, the rank distribution has 18% as Project Manager, 7% as System Analyst and 18% as Senior developers. Therefore, a 43% of the total manpower belongs to expert professionals. The working force consumes 46% share with Developers and Junior developers having 28% and 18%, respectively. This distribution matches the ideal manpower distribution that an established software company should have [26][27].

Alongside, the formation of the development team with respect to *number of people involved* in a project, adheres to the standard of agile practices [28]. Fig. 2(b) shows the typical formation of teams in the last 28 projects that are completed by these companies. As per this statistics, 9 projects had 5-9 people which is the standard for projects of medium duration, and 13 projects had either 10-15 people or 15+ people which is the conventional choice for large projects. Therefore, the companies are often guided by the standards when it comes to the matter of involving adequate manpower to the deserving projects. This is one of core concern in project management to ensure quality product development [30] [29].

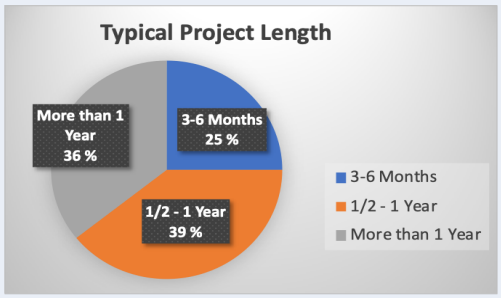
Finally, the selection and use of contemporary tools and techniques play a pivotal role in practicing agile methodologies and ensuring the quality product development. With the growing adoption of agile practices over the past couple of decades, a number of tools become the de-facto integral part of them. This includes, for instance, the version control systems (e.g. Git, GitHub), project management tools (e.g., Burn down charts, Jira) and project specific technologies (e.g., frameworks and languages), among others. The survey summary on this concern is shown in Fig. 4. Around 90% of the companies use Git as a version control system, and 60% of them use UML as a tool for technical design. Among the frameworks, web, ASP .net and app specific frameworks are used. This outcome is also inline with the project focus of the companies. However, professional project management and tracking tools are not used that frequently.

B. On the Agile Practice of the Software Companies

This research selected four Agile methods for the survey, namely, Scrum, XP, FDD and Crystal. These methods are selected based on their popularity in use. However, according to the survey response none of the companies ever used *Crystal method* for their projects, therefore, discarded from the discussion. Among the other three methods, Scrum is reported as the highest practiced method (82% of the companies use it) with XP and FDD having usage percentage of 4% each. Again due to very low response for XP and FDD, this research lacks sufficient empirical data to comprehensively assess the

Project Focus	Companies
Mobile Application	6
Banking Solutions	7
Management Solutions (HR, Project, Accounting)	9
E-Commerce & Web	10
AI, NLP, BIG data	2
Game Development	3
Others	5

(a)



(b)

Fig. 3. (a) Project Focus and (b) Typical Project Length of the Companies.

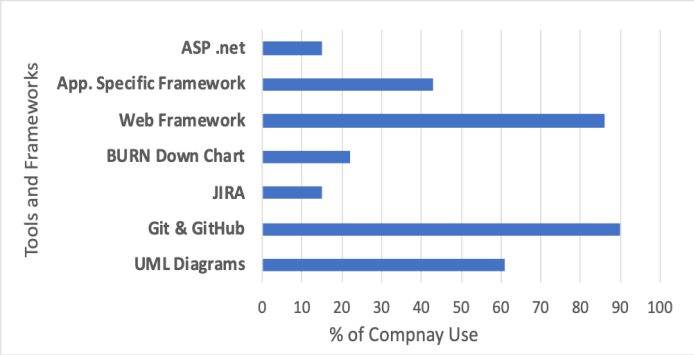


Fig. 4. Tools and Technologies used by the Companies.

adoption of these two methods. Therefore, exempted from further discussion.

In the following section a detail evaluation of the *Scrum method* is presented in relation to its' adoption and practice within the selected software companies. Additionally, acquired evidences and statistics are examined against the method to trace the followings, (a) the extent to which current practices resembles the standards, and (b) verify the preparedness of the companies in carrying out the projects by leveraging the method.

C. Scrum as a Development Method

Scrum as an agile method is the most popular development method according to this survey. 82% of the surveyed companies have adopted this method to carryout their development projects. To gain maximum insight on the topic, 16 questions in four distinct categories concerning the Scrum method are asked. These categories are, *Team (2 Questions), Artifacts (5 Questions), Role (2 Questions) and Process (6 Questions)*. Fig. 14 details this question set.

The ***Scrum Team*** should be assessed by their physical location and the team size for individual projects. The response on these concerns are highlighted in Fig. 5(a) and (b). As stated in the Scrum principles [5], the scrum team should be located in the same physical premises to maximize the effective communication among the team members for rapid development. However, in unavoidable circumstances, team members can be geographically distributed and collaborating over online. The survey reported that majority of the software firms are well within this recommendation (Fig. 5(a)). 53% of the companies always have on premises team with 31% sometimes. Only those companies that have off-shore sites have distributed teams (11%).

On the team size, Scrum practice suggested the standard should be seven, plus or minus two [5], having the range between 5 to 9 members. This number includes the *Scrum Master, Product owner* and the *Developers*. A team smaller than this recommendation may find it arduous to accomplish enough in each sprint, whereas for larger teams communication becomes complex and cumbersome [26][31]. The survey response (Fig. 5(b)) reported that only 48% of the companies maintain the recommended team size of 5 to 9 members, and the rest (52%) have either undersized or oversized team. Therefore, the companies must reassess their team formation with proper justification of performance and output produced.

Within the Scrum practices, two key responsibilities are to anchoring the *daily meeting* (a brief meeting held daily with the scrum team to synchronize development activities) and the *scrum review meeting* (a meeting held at the end of each sprint to assess the passing sprint and set goals for the next sprint) [32]. According to Scrum standard, the *Scrum Master* is the person who plays the ***Role*** of the anchor for these meetings. However, both the meetings (i.e. daily meeting and scrum review meeting) must be duly conducted by the *Team Members* [32]. Fig. 6(a) and (b) summarizes the survey outcome on this concern.

It is observed that in case of 70% of the companies, the *Scrum Master* is responsible for holding the daily review meeting whereas in 18% and 3% cases held by the team members and daily trackers, respectively (Fig. 6(a)). In case of Sprint review meeting, mostly Scrum master (53%) leads the meeting with 25% cases held by the team members and daily trackers (Fig. 6(b)). Therefore, it can be affirmed that in most part the companies adheres to the scrum mandate in maintaining the roles of the scrum team. Albeit, there are few companies who are involving the *Product owner* to anchor the designated meetings, which is neither desirable nor recommended by Scrum. Therefore, requires further explanation and rectification.

The Scrum method leveraged several means or ***Artifacts*** to

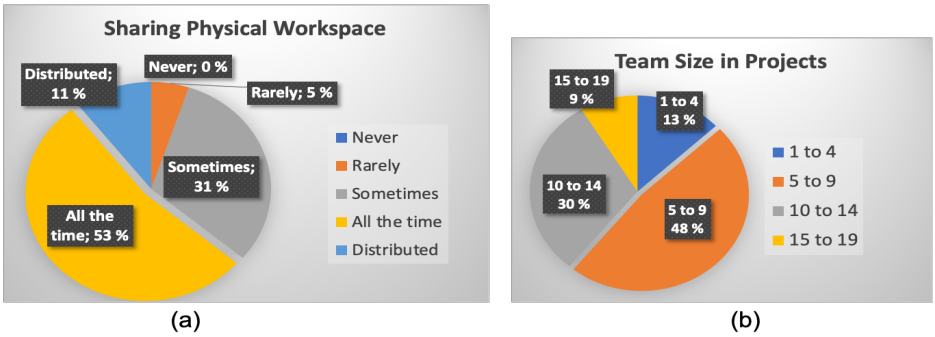


Fig. 5. Response on (a) Sharing Physical Space and (b) Project Team Size.

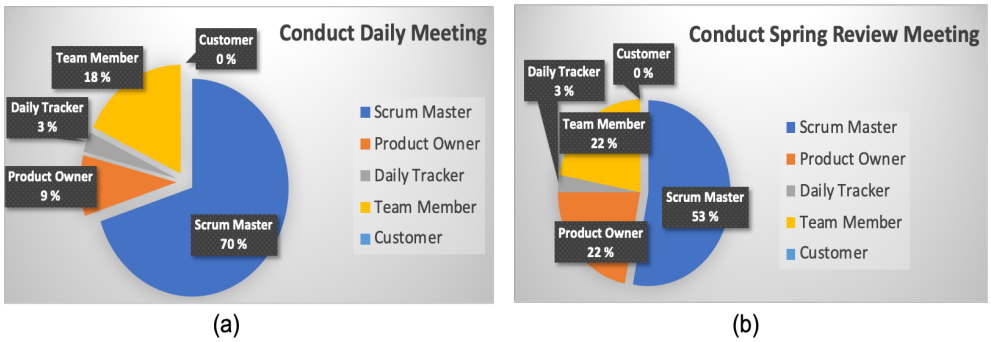


Fig. 6. The Team Member Responsible for Conducting (a) The Daily Meeting and (b) The Scrum Review Meeting

.

carryout the development activities. This includes for example, maintaining and following *Product Backlog, Sprint Backlog, Burndown chart*, among others. Companies practicing *Scrum* method should adopt and utilize these artifacts for efficacious product development [26] [5]. The survey outcome on this concern is summarized in Fig. 7(a), (b) and 8(a).

For reference, the *Product Backlog* describes the work to be done that will add value to the completed product. It is dynamic in nature to capture what are the most important features to be developed next. Therefore, the scrum master should constantly update and refine the Product Backlog to keep it aligned with market demand [32]. Whereas, the *Sprint Backlog* depicts the product increment to be implemented and added to the already done product at the end of current Sprint [31]. It should define two things: the "What to be developed" of the Sprint and the "How to develop" of the Sprint. It therefore, contains the blue print for the developers of how they will deliver the product Increment and realize the Sprint Goal [32].

The survey result on the use of backlog (either, Product or Sprint backlog) is detailed in Fig. 7(a). According to this reporting, about half of the companies (44%) use them for estimating the future requirements, 26% use them to prioritize the requirements and others, to record the requirements (13%) or to record the status (13%). However, according to practice, all these activities should be part of utilization of these backlogs [31].

The other core Scrum artefact is the *Burndown chart*, which is a graphical representation of work left to do over the project time [32]. This chart plots the outstanding work on the 'y-axis' with project time along the 'x-axis'. This visual representation helps the team to constantly monitor the project scope creep, and keep development work on schedule. This chart must be updated in the daily scrum meeting. However, the survey result on use of this chart differs largely with the proposals, as shown in Fig. 7(b). Only, 21% of the companies always use this chart (17%) or its' third-party variants (4%). Majority of them either rarely (70%) or never (9%) use it. Therefore, it is a major concern form Scrum perspective and the companies must put serious effort on adopting this tool as an integral part of their development practices.

In the realm of Scrum practices, using the above listed artefacts share the same goals. Those are, to maximize transparency through highly visible real-time picture of what is being done, and a shared understanding of the work in progress [26]. Therefore, these artifacts must be openly available to all the team members to see, discuss, follow and update to synchronize the rapid development activities [31]. The surveyed companies also adhere to this practice of making artefacts openly accessible to all the team members, either always (78%) or sometimes (22%) (Fig. 8(a)).

The *Scrum Process* defines the agile project management methodology for rapid development of a quality software product. This involves carrying out several activities by employing the team members in different roles and effective utilization of the artefacts. The core scrum activities includes, defining and updating the sprint backlog for a given sprint, holding the scrum meeting, sprint planning meeting and sprint review meeting, code integration and testing, and system demonstra-

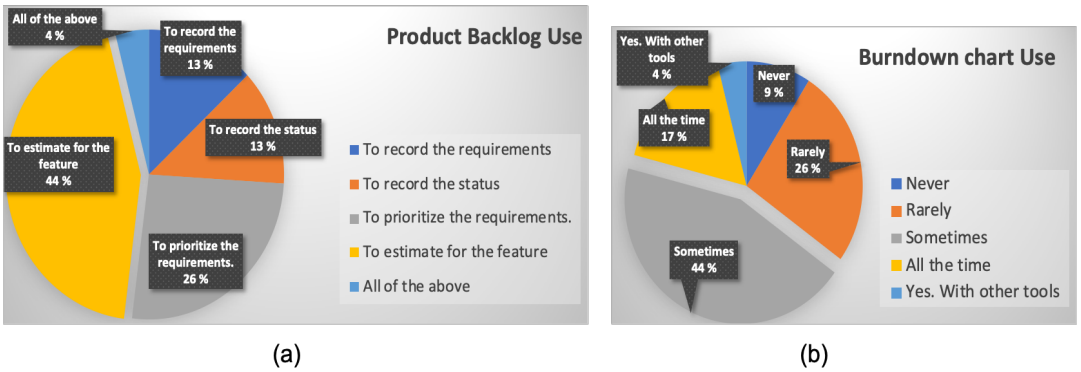


Fig. 7. Use of (a) Product Backlog and (b) Burndown Chart.

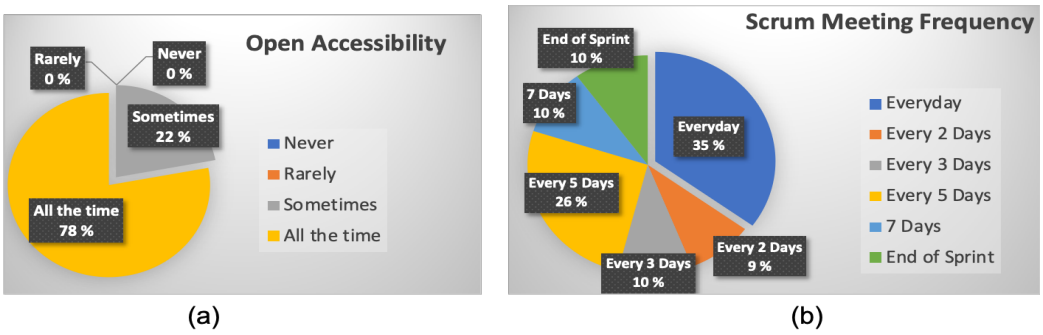


Fig. 8. (a) Accessibility of the Artifacts and (b) Scrum Meeting Frequency.

tion.

According to the methodology [31], a Scrum meeting is 15-minute time-boxed event that should be held each day of a sprint with the whole team. Inline with this recommendation, the survey response (Fig. 8(b)) exhibit that all the companies hold scrum meeting. However, the meeting is hold according to their own defined intervals, which varies from the daily meeting (35%), to holding it in every two days (9%), or in every 3 days (10%), or in 5 days (26%), 7 days (10%) and even at the end of sprint (10%). This statistics highly contradicts with the core value of agile practices and the scrum. Because, a sprint is usually lasts for 7 days with a sprint backlog to be implemented. Therefore, holding scrum meeting daily is an inevitable need for the development to progress smoothly. However, 65% of the companies are not realizing the fact, and therefore, suffers from absorbing the core essence of scrum. Consequently, this reporting calls for further investigation and rectification in the process.

The *Sprint backlog* for a given sprint consists of a list of tasks selected from product backlog to be completed within the sprint [32]. As the sprint length is short and development goes rapid, the Sprint backlog should be updated *once each day* by the *Scrum Master* and the burndown chart is updated to keep every team member in sync [31]. Adoption of these practices within the surveyed companies are shown in Fig. 9(a) and (b).

Reporting on the Sprint backlog update frequency (Fig. 9(a)) reveals three distinct trends, namely, daily (only 40% of the companies adopt this), between 2 to 7 days (36% of the companies follow this), and only with client requirement change (22% of the companies). Therefore, companies have to revise their understanding and practice on this particular concern. However, in 88% of the companies either the Scrum Master (40%) or a designated team member (40%) is responsible to update the backlog (Fig. 9(b)), which is well within the scrum convention.

Among the other core tasks, conducting Sprint Planning and Sprint review meetings with the involvement of the *Product owner* is highly recommended. Fig. 10 and 11 details the survey outcome on these practices. According to the Scrum guide, the *Sprint Planning meeting* is held at the beginning of each sprint to set the sprint backlog. The *Sprint Review meeting* is held at the end of a Sprint to inspect whether the backlog is implemented accordingly. Among the other stakeholders, the *Product Owner* must be present in the meetings to prioritize the most important features to be implemented and verified [26].

Majority of the companies (61%) agrees that they always hold the sprint planning meeting with 34% respond with either sometimes or rarely (Fig. 10(a)). In defining the length (or duration) of a Sprint, 48% companies maintains the highly recommended 7 days window, whereas, 38% responds with either 14 or 30 days duration and 14% says its' depends on the project. At large, neither of these statistics follows the recommendation, and is a violation of the core practices of Scrum methodology.

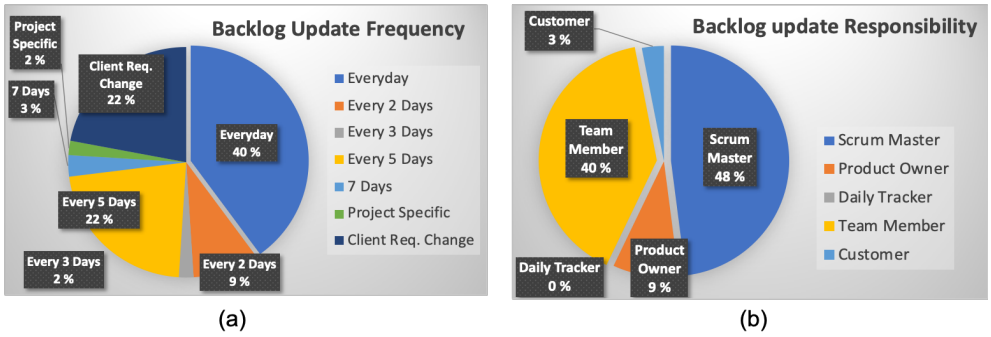


Fig. 9. (a) Sprint Backlog update Frequency and (b) Person Responsible to do the Update.

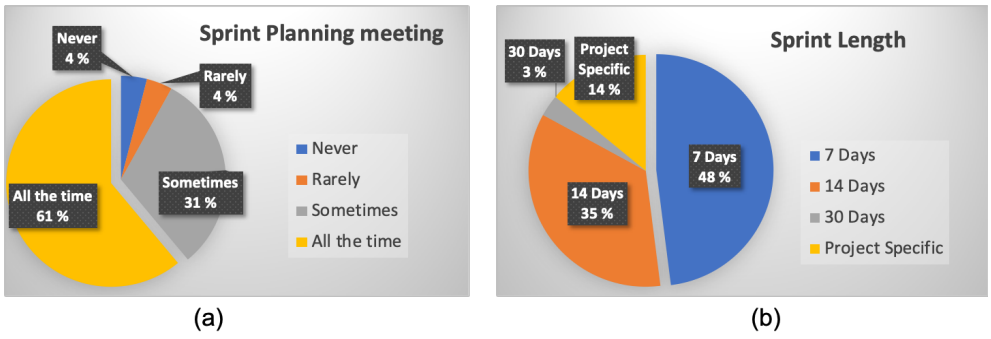


Fig. 10. (a) Holding of Sprint Planning Meeting and (b) Typical Length of a Sprint.

However, 70% of the companies responded positively in relation to hold the Sprint review meeting always (Fig. 11(a)), with 30% either sometimes or rarely holding the meeting. Therefore, companies are better performing in relation to this core activity of scrum. The Product owner is rarely attending either of the meetings according to the survey outcome (Fig. 11(b)). With the fact that the attendance of Product owner is highly recommended in the sprint meetings, only 9% of companies acknowledge their presence all the time. For the rest (91%) it is either sometimes, rarely or never. This outcome also point to the fact that the product owner in Bangladesh might lag the technical competencies or the client companies are reluctant to involve their representatives to cut cost. Whatever may be the reason, this lagging in participation is detrimental to overall process adoption and to the quality of the software produced [32].

Scrum methodology like other agile practices relies on continuous code integration on the daily basis [31]. Integration testing must go hand-in-hand with the daily integration [31]. However, the survey outcome shows a large deviation with this standard practice. As can be seen from Fig. 12(a), only 44% companies adheres to daily integration and testing, while majority have their own defined schedule.

Finally, developed system (either at the end of each sprint or at the end of the project) is demonstrated practically by executing it [32], rather using any means of formal presentations (e.g., power points, oral or visualization). According to the survey response (Fig. 12(b)), majority (57%) follows the convention of demonstrating the system practically, while others use undesirable methods.

V. Overall Assessment

The overarching assessment of the survey outcome highlights both competencies and weaknesses of the software companies in relation to Agile fitness. The taxonomy of the company portfolio reveals strong competencies to adopt and practice agile development methods (as discussed in detail in Section IV-A). The overall operational experience, range of software product development and service delivery expertise, the formation of the development team, the selection and use of contemporary tools and techniques, strongly support this claim. Therefore, it can be affirmed that

> *To a great extent the software companies in Bangladesh have the readiness to effectively practice Scrum methodology.*

Among the Agile methodologies, the *Scrum method* has overwhelming utilization in the software companies (82%) in Bangladesh. This selection reflects the most prevalent choice worldwide, as 70% software companies goes by the Scrum method [25]. However, the critical assessment of the survey statistics on the actual adoption of Scrum practices (as detailed in Section IV-B) reveals that

> *At large, the software companies in Bangladesh fall short to comply with the Scrum principles.*

Fig. 13 summarises the Scrum fitness of the companies derived from the survey results. In this figure, the approval / adoption rate (in X-axis) of companies are shown against the recommended Scrum practices (in Y-axis).

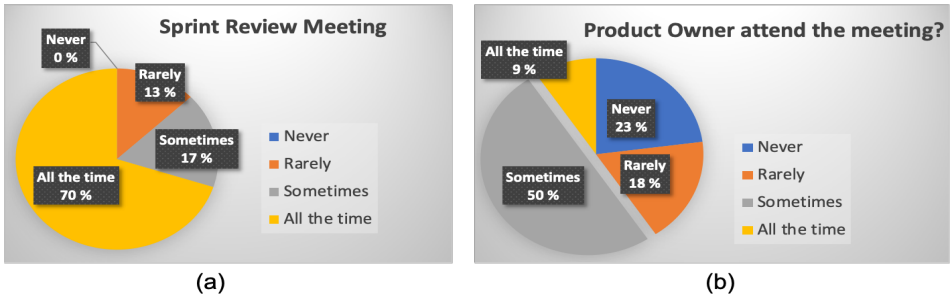


Fig. 11. (a) Holding Sprint Review Meeting and (b) Attendance of Product Owner in the Sprint Planning and Review Meetings.

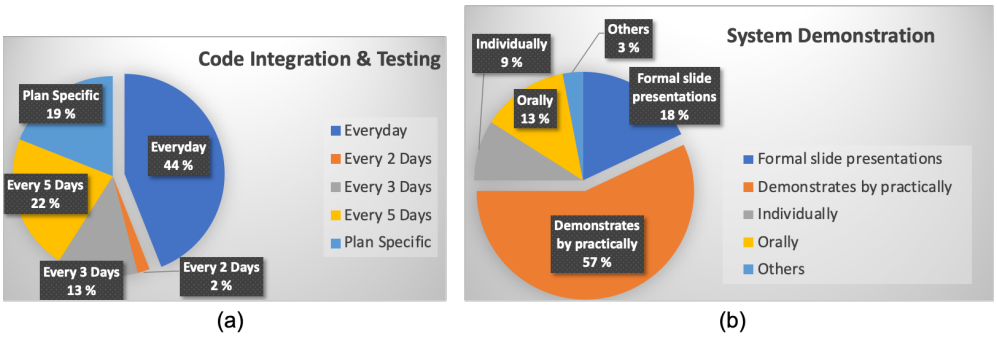


Fig. 12. (a) Code Integration and Testing Frequency, (b) Method used for System Demonstration.

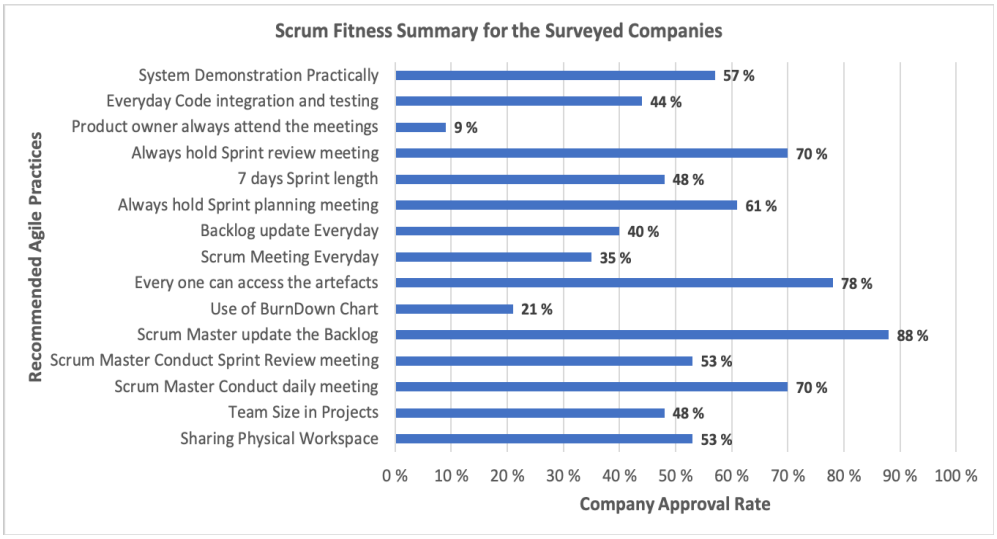


Fig. 13. Scrum Fitness overview for the Surveyed Companies.

According to the summary report in Fig. 13, the approval rate is around 50% or bellow for most of the key practices of Scrum. For some practices the rate is critically low which is alarming. For instance, integrating the product owner in the meetings is only 9% which on the contrary is one of the highest priority practices to be adopted [32]. Additionally, use of burndown charts or similar tools for constantly tracking the project progress and keep all the stakeholders in synchronized is only 21%. For the other practices along with the above two, the adoption rate need to be improved. This study recommences the companies to employ agile experts external to the company to investigate into the issues, identify core areas of improvement and a pragmatic course of actions to meet the Scrum standard [31].

VI. Conclusion

This research carried out an empirically investigation on the agile software development practices within the context of established software companies in Bangladesh to (a) define the readiness and fitness of these companies in relation to Agile practices, and (b) formulate the scope of improvements based on the agile standard. It is reported that the Scrum agile

method is the highest practiced one among the four, which is an assertion of the typical selection worldwide. Alongside, the overarching outcome reveals that the companies have the preparedness in practicing the Scrum method in fullest. However, with regard to practicing Scrum principles, they fall short severely in many key factors. Therefore, the future research should dug deep into the cause of these shortcomings and formulate guidelines accordingly for the process improvement.

SCRUM METHOD QUESTIONNAIRE

Scrum Questionnaire Set (Fig. 14).

REFERENCES

[1] Eva-Maria Schön, Jörg Thomaschewski, María José Escalona, Agile Requirements Engineering: A systematic literature review, Computer Standards & Interfaces, Volume 49, 2017, Pages 79-91.

[2] E.M. Schön, M.Escalona, J.Thomaschewski, Agile Values and Their Implementation in Practice,International Journal of Interactive Multimedia and Artificial Intelligence, 3 (61), 2015.

[3] Misra, S., Kumar, V., Kumar, U., Fantazy, K. and Akhter, M. (2012), Agile software development practices: evolution, principles, and criticisms, International Journal of Quality & Reliability Management, Vol. 29 No. 9, pp. 972-980.

[4] C. Dewan, R. Jain, and R. Kohli, The Agile Methodology, IJCSMS International Journal of Computer Science & Management Studies, 12(3), September 2012.

[5] K. Schwaber, M. Beedle, Agile Software Development with Scrum, vol. 1, Prentice Hall Upper Saddle River, 2002.

[6] K. Beck, C. Andres, Extreme Programming Explained: Embrace Change, Addison-Wesley Professional, 2004.

[7] M. Poppendieck, T. Poppendieck, Lean Software Development: An Agile Toolkit, Addison-Wesley Professional, 2003.

[8] D.J. Anderson, Kanban, Blue Hole Press, 2010.

[9] P. Rodríguez, J. Markkula, M. Oivo, K. Turula, Survey on Agile and Lean usage in finnish software industry, in: Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM '12, ACM, New York, NY, USA, 2012, pp. 139–148.

[10] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R.C. Martin, S. Mellor, K. Schwaber, J. Sutherland, D. Thomas, Manifesto for Agile Software Development, 2007.

[11] Eetu Kupiainen, Mika V. Mäntylä, Juha Itkonen, Using metrics in Agile and Lean Software Development – A systematic literature review of industrial studies, Information and Software Technology, Volume 62, 2015, Pages 143-163.

[12] S.R. Palmer,M.Felsing, A Practical Guide to Feature-Driven Development, Pearson Education,2001.

[13] D. Duka, Adoption of agile methodology in software development, Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013.

[14] Mahanti A. Challenges in Enterprise Adoption of Agile Methods - A Survey. Journal of Computing and Information Technology - CIT 14, 2006, 3, 197–206.

[15] . Qumer, B. Henderson-Sellers, A framework to support the evaluation, adoption and improvement of agile methods in practice, Journal of Systems and Software, 81(11), 2008, Pages 1899-1919.

[16] Gurpreet Singh Matharu, Anju Mishra, Harmeet Singh, and Priyanka Upadhyay. 2015. Empirical Study of Agile Software Development Methodologies: A Comparative Analysis. SIGSOFT Softw. Eng. Notes 40, 1 (January 2015), 1–6.

[17] Nageswara KudaPavan G PPavan G PNaidu KavitaPraneeth Chakka, A Study of the Agile Software Development Methods, Applicability and Implications in Industry, International Journal of Software Engineering and its Applications 5(2), 2011.

[18] J.F. Tripp, D.J. Armstrong. Exploring the Relationship Between Organizational Adoption Motives and the Tailoring of Agile Methods, Hawaii International Conference on System Science, pp. 4799-4806, 2014.

[19] Stapleton, J. DSDM: Dynamic Systems Development Method. Addison, Reading, MA, 1997.

[20] ICT Business Promotion Council Ministry of Commerce, Peoples Republic of Bangladesh, Visited in 2020. http://www.bpc.org.bd/ibpc_software_industry.php.

[21] Kumkum Katha, Software Company in Bangladesh Contributing to vision 2021. Visited in 2020. https://www.southtechgroup.com/software-company-in-bangladesh/.

[22] Jon A. Krosnick, Survey Research, Annual Review of Psychology, Vol. 50, pp. 537-567. 1999. https://doi.org/10.1146/annurev.psych.50.1.537.

[23] Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In H. T. Reis & C. M. Judd (Eds.), Handbook of research methods in social and personality psychology (p. 223–252). Cambridge University Press.

[24] Vartika Kashyap, Best Business Management Software You Should be Using Today. 2020. https://www.proofhub.com/articles/best-business-management-software.

[25] Murat Yilmaz, Rory V. O'Connor, and Paul Clarke. 2015. Software Development Roles: A Multi-Project Empirical Investigation. SIGSOFT Softw. Eng. Notes 40, 1 (January 2015), 1–5.

[26] R. Hoda, J. Noble and S. Marshall, "Self-Organizing Roles on Agile Software Development Teams," in IEEE Transactions on Software Engineering, vol. 39, no. 3, pp. 422-444, March 2013, doi: 10.1109/TSE.2012.30.

[27] Baddoo, N. and Hall, T. (2002), Practitioner roles in software process improvement: an analysis using grid technique. Softw. Process: Improve. Pract., 7: 17-31. doi:10.1002/spip.151.

[28] V. Lalsing, S. Kishnah, S. Pudaruth, People Factors in Agile Software Development and Project Management, Journal of Software Engineering & Applications (IJSEA), Vol.3, No.1, January 2012 DOI : 10.5121/ijsea.2012.3109.

[29] S. W. Ambler, "Scaling agile software development through lean governance," 2009 ICSE Workshop on Software Development Governance, Vancouver, BC, 2009, pp. 1-2, doi: 10.1109/SDG.2009.5071328.

[30] A. Ahmed, S. Ahmad, N. Ehsan, E. Mirza and S. Z. Sarwar, "Agile software development: Impact on productivity and quality," 2010 IEEE International Conference on Management of Innovation & Technology, Singapore, 2010, pp. 287-291, doi: 10.1109/ICMIT.2010.5492703.

[31] C. Larman, Agile and Iterative Development: A Manager's Guide. Boston: Addison Wesley, 2004.

[32] K. Schwaber, "SCRUM Development Process," Business Object Design and Implementation, pp. 117–134, 1997.

Agile Method Concern	Questions	Answer Options
The Team *Working place*	Do all the team members work in the same space?	1. Never 2. Rarely 3. Sometimes 4. All the time 5. Other .....
Team size	What is the overall size of a scrum team?	1. 1-4 2. 5-9 3. 10-14 4. 15-19 5. Other .....
Artifacts *Artifacts visibility*	Are the artifacts (Product Backlog, Sprint Backlog, Burndown chart) openly accessible and visible to the Scrum Team?	1. Never 2. Rarely 3. Sometimes 4. All the time 5. Other .....
Sprint Burndown Chart	Are Sprint burndown charts used to track the progress of the project?	
Product Backlog	What is the purpose of the Product backlog in your institution?	1. To record the requirements 2. To record the status 3. To prioritize the requirements. 4. To estimate for the feature 5. All of the above
Sprint Backlog	How frequently is the Sprint backlog updated?	1. Everyday 2. Every 2 days 3. Every 3 days 4. Every 5 days 5. Other .....
Updated by	Who updates the Sprint backlog?	1. Scrum master 2. Product Owner 3. Daily Tracker 4. Team member 5. Customer
Role *Daily Meeting*	Who conducts daily Scrum meetings?	
Conducted by	Who conducts the sprint review meeting?	
Process *Sprint planning meeting*	In your institution, are Sprint planning meetings conducted? Are Product owners and representatives present in the Sprint Planning Meeting?	1. Never 2. Rarely 3. Sometimes 4. All the time 5. Other .....
Sprint review	At the end of each Sprint, is there any Sprint review to discuss the progress of the project?	
Sprint length	What is the general length of each Sprint?	1. 7 days. 2. 14 days. 3. 30 days. 4. Over 30 days. 5. Other .....
Scrum meeting	How frequently are Daily Scrum meetings held?	1. Everyday 2. Every 2 days 3. Every 3 days 4. Every 5 days 5. Other .....
Testing the code	How frequently is the code integrated and tested?	
Demonstration process	How is the developed system demonstrated?	1. Formal slide presentations 2. Demonstrates by practically 3. Individually 4. Orally 5. Other .....

Fig. 14. Scrum Method: Questionnaire with Answer Options