



KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

ISSN (P): 2616-9592

ISSN (E): 2710-5075



Volume: 5

Issue: 2

Jul - Dec

2022



KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

Volume 5, Issue 2, 2022

ISSN (P): 2616-9592

ISSN (E): 2710-5075

Frequency Bi-Annual

Editorial Board

Patron

Air Vice Marshal (Retd) Tubrez Asif, HI(M) - President, KIET

Editor-in-Chief

Prof. Dr. Muzaffar Mahmood

Associate Editor

Dr. Muhammad Affan Alim

Managing Editor

Prof. Dr. Muhammad Khalid Khan

Manager Production & Circulation

Syed Hassan Ali

College of Computing & Information Sciences
Karachi Institute of Economics & Technology

College of Computing & Information Sciences

Vision

To develop technology entrepreneurs & leaders for national & international market

Mission

To produce quality professionals by using diverse learning methodologies, aspiring faculty, innovative curriculum and cutting edge research, in the field of computing & information sciences.



AIMS AND SCOPE

KIET Journal of Computing and Information Sciences (KJCIS) is the bi-annual, multi-disciplinary research journal published by **College of Computing & Information Sciences (CoCIS)** at **Karachi Institute of Economics and Technology (KIET)**, Karachi, Pakistan. **KJCIS** aims to provide a panoramic view of the state of the art development in the field of computing and information sciences at global level.

It provides a premier interdisciplinary platform to researchers, scientists and practitioners from the field of computing and information sciences to share their findings and contribute to the knowledge domain at global level. The journal also fills the gap between academician and industrial research community.

KJCIS focused areas for publication includes; but not limited to:

- Data mining
- Big data
- Machine learning
- Artificial intelligence
- Mobile applications
- Computer networks
- Cryptography and information security
- Mobile and wireless communication
- Adhoc and body area networks
- Software engineering
- Speech and pattern recognition
- Evolutionary computation
- Semantic web and its application
- Data base technologies and its applications
- Internet of things (IoT)
- Computer vision
- Distributed computing
- Grid and cloud computing

OPEN ACCESS POLICY

For the benefit of authors and research community, this journal adopts open access policy, which means that the authors can self-archive their published articles on their own website or their institutional repositories. The readers can download or reuse any article free of charge for research, further study or any other non profitable academic activity.

PEER REVIEW POLICY

Peer review is the process to uphold the quality and validity of the published articles. KJCIS uses double-blind peer review policy to ensure only high-quality publications are selected for the journal. Papers are referred to at least two experts as suggested by the editorial board. All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. We expect our Board of Reviewing Editors and reviewers to treat manuscripts as confidential material. The identities of authors and reviewers remain confidential throughout the process.

COPYRIGHT

All rights reserved. No part of this publication may be produced, translated or stored in a retrieval system or transmitted in any form or by any means; electronic, mechanical, photocopying and/ or otherwise the prior permission of publication authorities.

DISCLAIMER

The opinions expressed in **KIET Journal of Computing and Information Sciences (KJCIS)** are those of the authors and contributors, and do not necessarily reflect those of the journal management, advisory board and the editorial board. Papers published in KJCIS are processed through double blind peer-review by subject specialists and language experts. Neither the **CoCIS** nor the editors of **KJCIS** can be held responsible for errors or any consequences arising from the use of information contained in this journal, instead; errors should be reported directly to the corresponding authors of the articles.

Academic Editorial Board

Dr. Ronald Jabangwe University of Southern Denmark, Denmark	Dr. Sardar Anisul Haque Alcorn State University, USA
Dr. M. Ajmal Khan Ohio Northern University, USA	Dr. Yasser Ismail Southern University Louisiana, USA
Dr. Suliman A. Alsuhibany Qassim University, Saudi Arabia	Dr. Manzoor Ahmed Hashmani University of Technology Petronas, Malaysia
Dr. Wael M El-Medany University of Bahrain, Bahrain	Dr. Atif Tahir FAST NUCES, Pakistan
Dr. Asim Imdad Wagan Mohammad Ali Jinnah University, Pakistan	Dr. Maaz Bin Ahmed Karachi Institute of Economics & Tech, Pakistan
Dr. Salman A. Khan Karachi Institute of Economics & Tech, Pakistan	Dr. Taha Jilani Karachi Institute of Economics & Tech, Pakistan

Advisory Board

Dr. Andries Engel brecht University of Pretoria, South Africa	Dr. Mohamed Amin Embi University Kebangsaan, Malaysia
Dr. Rashid Mehmood King Abdul Aziz University, Saudi Arabia	Dr. Anh Nguyen-Duc Norwegian University of Technology, Norway
Dr. Ibrahima Faye University of Technology Petronas, Malaysia	Dr. Tahir Riaz Data Architect, SleeknoteApS, Denmark
Dr. Faraz Rasheed Microsoft, USA	Dr. Mostafa Abd-El-Barr Kuwait University, Kuwait
Dr. Abdul Naser Mohamed Rashid Qassim University, Saudi Arabia	Dr. Mohd Fadzil Bin Hassan University of Technology Petronas, Malaysia
Dr. Syed Irfan Hyder Ziauddin University, Pakistan	Dr. Bawani S. Chowdry Mehran University, Jamshoro, Pakistan
Dr. Jawad Shami FAST - NUCES, Pakistan	Dr. Nasir Tauheed Institute of Business Administration, Pakistan

Table of Content

1 01-20	An Intuitive Architecture for DIY IoT Application Composition as Business Process Model <i>Muhammad Sohail Khan, DoHyeun Kim and Faiza Tila</i>
	The Impact of Mitigation Strategies on Geographical Distance Issues in GSD: An Empirical Evaluation <i>Nadia Ka e nat, Uzair Iqbal Janjua, Tahir Mustafa Madni and Atta ur Rahman</i>
2 21-40	
3 41-66	Comparative Analysis of Machine Learning techniques to Improve Software Defect Prediction <i>Muhammad Azam, Muhammad Nouman and Ahsan Rehman Gill</i>
	Statistical Analysis for the Traffic Police Activity: Nashville, Tennessee, USA <i>M. Y. Tufail and S. Gul</i>
4 67-84	
5 85-100	An Assessment for Understanding Student Behaviour by Applying Machine Learning Technique <i>Syeda Kainat Ahmed, Syed Mushhad M. Gilani, Sidra Sultan, Abdur Rehman Riaz and Muhammad Wasim Abbas</i>
	Enhanced Accessibility of Facebook Messenger for Blind Users <i>Mamoona Atif Swati, Dr. Mustafa Madni, Dr. Uzair Iqbal Janjua and Dr. Iftikhar Ahmed Khan</i>
6 101-116	
7 117-126	Automatic Speech Recognition on Non-Pathological Dataset of Urdu Language <i>Anoshia Imtiaz, Munaf Rashid, Sidra Abid Syed, Hira Zahid, Muzaffar Iqbal and Akhtar Ali Khan</i>

An Intuitive Architecture for DIY IoT Application Composition as Business Process Model

Muhammad Sohail Khan¹

DoHyeun Kim²

Faiza Tila²

Abstract

Internet of Things (IoT) and the related technologies have changed the users' perspective of remote service utilization. Internet of Things is a network of billions of smart and connected devices, which expose their functionality in the form of sensing or actuation services. Currently, applications utilize those services to cater to the user's needs. However, a user's requirements are always changing which cannot be efficiently fulfilled by the domain specific applications. We propose the idea of using Business Process Modeling approach in order to enable the user to model their required processes based on the atomic services exposed by IoT devices. The approach requires no programming skills on behalf of the users so it can be used as a Do-It-Yourself (DIY) tool for the creation of IoT based processes. This article presents the architecture and detailed design of the proposed system.

Keyword: Business process modeling, Internet of Things, Architecture, DIY, Application.

1. Introduction

Recent years have shown the growth in business and research related to Internet of Things (IoT). According to IoT Analytics, by the year 2025, there will be 27.1 billion devices connected to the internet while currently it is around 13 billion, an unexpectedly low growth due to the COVID-19 pandemic and chip shortage [1]. The International Data Corporation (IDC) reported in its latest reports, that the IoT spending in the Asia Pacific region only will reach 437 billion USD till 2025 with an expansion of 9.6% as compared to the 2020 growth rate of 1.5% [21]. This encourages the academia and industry to strive for the realization of IoT vision. Several standardization efforts are underway and IoT architectures have been presented with a focus to standardize the way IoT systems and applications are developed.

Business Process Management (BPM) has been at the center of close collaboration between business and IT for a long time. This popularity of the BPM solutions is mostly due to the standardization of diagramming language known as the Business Process Modeling Notations (BPMN) [2]. The initial aim of BPMN was to enable the business

¹University of Engineering & Technology, Mardan, Pakistan | sohail.khan@uetmardan.edu.pk

²Jeju National University, South Korea | kimdh@jejunu.ac.kr, faizakhan797@gmail.com

analyst to describe a business's desired process through a diagram and to accommodate the business agility through automated execution of the process model.

Service Oriented Architecture (SOA) [3] has since been at the forefronts of BPM solution implementations with the promise of service reuse. Service reuse is the encapsulation of a system's atomic functions as reusable service units with well-defined interfaces. These reusable services can provide easy and rapid integration of new composite processes hence making the IT to become more agile. BPMN can be utilized for more than just a means for business requirements gathering rather it is considered as a prominent solution for fast passed Industry 4.0 related process driven applications [4]. A recent survey by BpTrends [5] reveals that about 75 percent of the enterprises believe that BPM processes and technologies have helped them achieve their business goals, 73 percent reported an increasing interest in BPM since 2019 and 71 percent of the survey respondents reported the current digital transformation in the technology as the motivator towards BPM adoption. This indicates the importance and awareness of BPMN among businesses and its importance in the current technological scenario.

The current technologies in the form of sensors and actuators networks, web of things [6] and most of all the realization of the Internet of Things (IoT)[7] involves ever changing requirements and implementation within the ecosystem of resource constrained hardware, services and people. Until recently, accommodation of user desired change in applications associated with these paradigms was not easily possible due to the resource constraints, heterogeneous nature of the hardware and the lack of standardization in the communication strategies. With the recent improvements of security in REST [8], resource constrained protocols such as MQTT [9] and the recent CoAP [9] Protocol, a flexible service orientation has become possible for the things associated with IoT. Service composition and orchestration has been introduced to the recent developments in IoT and other associated paradigms. SENSEI [10] is a business driven IoT architecture for the scalability of large numbers of sensors and actuators associated with Internet of Things. It is focused on providing services for accurate retrieval of contextual information and interaction with physical entities.

MoCoSo[11] is another such project which is focused on the combination of various concepts such as identification of objects, contextual data and communication medium (Internet). The main idea is to integrate sensor networks into a larger Internet of Things. Systems focused towards the application of IoT in industry have also been initiated. Collaborative Business Items (CoBIs) project is a core project which aimed at enabling industrial objects such as machine parts and containers etc. to communicate with each other at their surroundings. This goal is achieved by making the items uniquely identifiable in the system and incorporating various sensors in the said items. The items exposed their behavior as services and the system provided an infrastructure for centralized service

deployment across the network.

Most of the above projects belong to the initial era of effort towards the realization of IoT vision and are at least 5 to 10 years old. A more recent project is the Compose project [20]. This project provides an open and scalable platform infrastructure enabling easy creation of application based on IoT. The main theme of the project is to represent the Internet connected smart objects as programming entities which can be utilized to program larger applications. This project, however, is composed of several open source technologies which form a complex integrated infrastructure in order to achieve its goals and in there somehow lacks the intuitiveness for daily life, non-programmer users that a DIY type of system can provide.

Research community has embraced the potentials of a resource constrained device in the paradigm of business process management and hence efforts have been done to include things and services as part of the BPMN. In this regard, [12] provided the missing concept of “thing”, as presented by the main components of IoT reference model [13], by extending the conventional meta-models for business process modeling notations. Similarly, [14] proposes the extension of the business process modeling lifecycle for the integration of IoT in it.

Despite the efforts to integrate IoT as part of the BPM lifecycle. Traditionally, business process modeling and the demand for rapid incorporation of change have always been based on service reuse and service composition. This practice, although proven effective at times, has not been always effective [4]. IoT is also not just some enterprise implementation of proprietary services for specific goals which can be composed upon to execute processes. In fact, IoT is a vast ecosystem of billions of devices which present themselves as atomic services on the Internet. These services must be available for masses to utilize for making their local solutions as well as to share them. This concept is also termed as the Do-It-Yourself (DIY) paradigm of development for the realization of IoT vision.

This idea has been advocated by many such as [15], [16] states that the end-users should be part of the creation process while having the power to discover things. Similarly, [16], [17] suggests that the end-users should be able to discover things and control them in order to effectively use the application for smart environments. The vision and motivations of Makers Revolution [18], the advancement in DIY prototyping platforms such as Arduino and Raspberry Pi etc. [19] and the ongoing standardization of communication protocols for constrained devices are all the right steps towards inculcating DIY culture in masses. However, the masses may not have the skills and the ability to program these embedded devices for their DIY implementations especially with the growing number of programming languages currently being used for IoT implementations. BPM and the associated diagramming language may prove useful as a solution to the problem.

BPM has always been envisioned as the tool to enable the managers and people with no programming skills to describe their needs and desired processes. We propose the utilization of BPM diagramming language for the IoT users to make their own desired processes and let the SOA enable those models to be executed in their environment. Our vision is to let the user model the process based on the available atomic services which have some level of composition and then integrate them with user defined rules to model their processes and directly execute those processes without the need to alter the underlying services according to the process model. Such an approach can enable the users to easily model and execute their desired processes and hence make IoT applications agile with respect to the user requirements.

This paper presents the architecture and design of the layered system for the realization of our vision. The rest of the paper presents a detailed description of the system architecture and design with the help of static and sequence models while the preliminary semantic model has been presented in the form of Protégé based implementation.

2 System Architecture

Fig. 1 shows the conceptual architecture for the system. The architecture consists of four layers each of which performs its own functions and communicates with the adjacent layers through a predefined communication scheme. A brief introduction of the system based on each layer is provided in the following paragraphs.

The physical layer consists of things associated with the Internet of Things. These things are capable of sensing data from their surroundings and/or controlling certain phenomenon happening around them. These things can simply be termed as sensing and actuating devices with the capability of communication through the internet.

The Virtual Object Layer (VOL) represents the physical layer things in the form of Virtual Objects. A Virtual Object (VO) is the in-system representation of a thing at physical layer. A virtual object encapsulates the information associated with a physical thing and enables the users to manipulate the VO inside the system environment, interact with it and through the VO interact with the environment of the physical things represented by the VO.

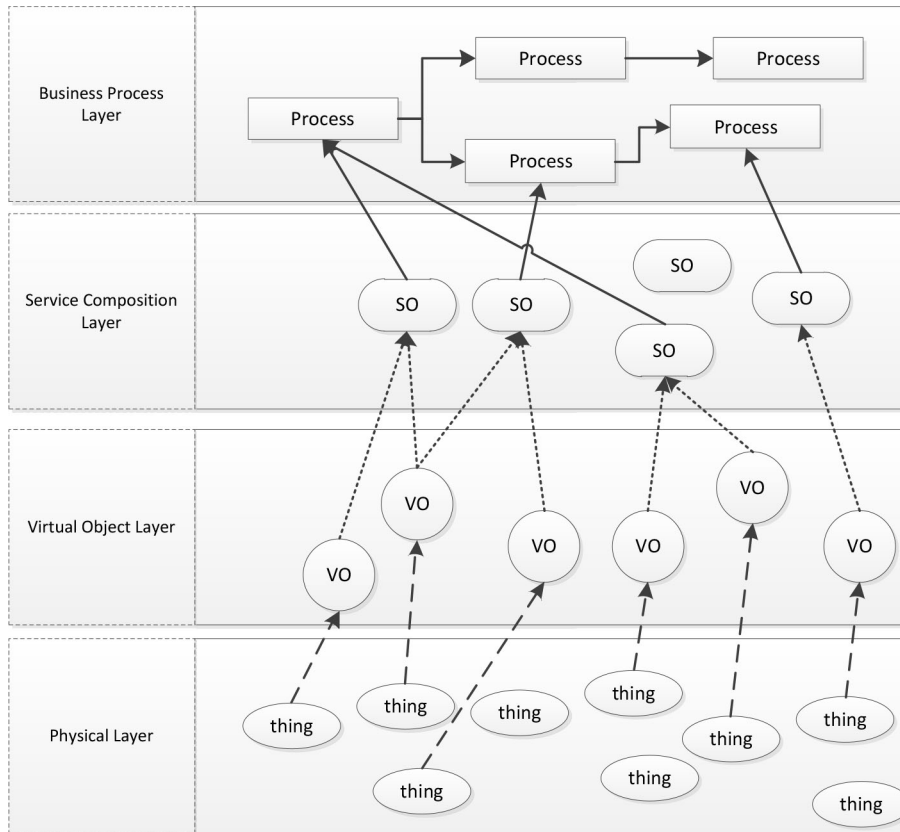


Figure 1: BPM based IoT system model

The virtual objects at the virtual object layer are utilized by the Service Composition Layer (SCL) in order to compose Service Objects (SO) by combining the functionalities offered by two or more virtual objects. A simple SO may thus contain an input VO joined with an output stream to display the data generated by the input VO. A more complex example of a service object would be to join a temperature sensor VO with an LED VO with the settings that when temperature value exceeds 40 degrees Celsius, the LED should start blinking. The acquisition of the temperature value and the blinking of the LED depend on the functionalities encapsulated by their corresponding VOs.

Once the SOs are created, these atomic service objects are utilized by the Business Process Layer (BPL) to formulate the flow of a user-desired process for a certain context. The business process layer utilizes the business process modeling notations (BPMN) to model the process based on the combination of SOs. The process flows are executed at the business process layer to carry out the functionality as defined by the individual service objects and actual interactions are carried out directly with the physical things based on their behavior encapsulated by their respective VOs.

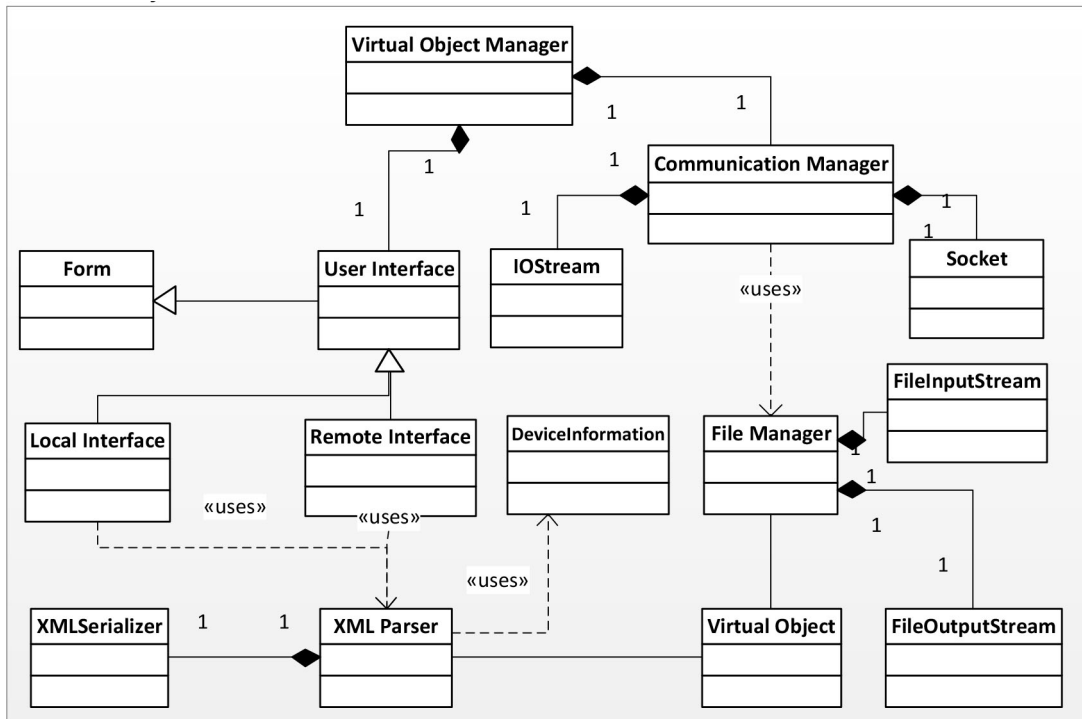


Figure 2: Static structure for virtual object manager

3. Detailed Design

This section presents the design details of the layers as presented in the previous section. The design of each layer includes the static structures and the interaction design for describing the main operations at each layer. The following sub-sections present the design details of each layer.

3.1 Object virtualization

The virtual object manager is the main component at the virtual object layer. It in collaboration with other classes such as the File Manager, Communication Manager and Parser etc., provides the implementation of all the functionality associated with the VOL. The static structure of VOM is shown in Fig. 2. VOM is the composition of the local and remote interfaces classes which enables the users to input information related to the physical things for which they want to register virtual objects. Similarly, the Communication Manager uses the File Manager for retrieving the XML version of the virtual objects from the local file system and to send it the client application. The client application in this scenario would be the service composition manager. The XML Parser works in collaboration with the File Manager and the interfaces to convert the information entered by users into xml elements representing the VO and vice versa. The

XML Parser uses the DeviceInformation class as a template for the creation of VOs.

Fig. 3 shows the internal process of the virtual object manager in the form of a sequence diagram. The sequence model shows the interaction of user with the interface component as well as the resultant interactions in the form of messages exchange among the other internal components of the system in order to fulfill the user commands. The sequence of interactions starts when the VOM is started and all the components including the user interface and the Communication Manager etc. are initialized.

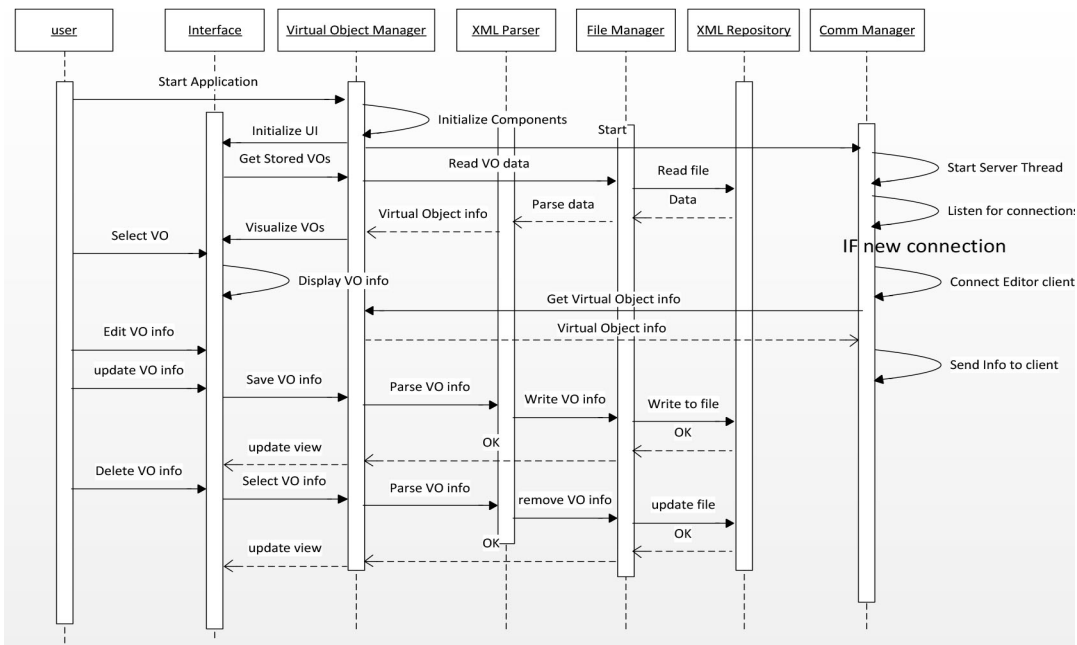


Figure 3: Virtual object manager operation sequence

The main interface provides a view for all the existing VOs so it requests the File Manager through the VOM to read the VO data from the XML repository. The data is parsed by the XML parser and the virtual object information is provided to the VOM in order to display it through the interface. Now the user is set to interact with the VOs through the interface. The VO related interactions that user can perform have been shown in the sequence model. The user selects a VO graphical representation and the VO information is displayed to the user through the view interface. The user can then choose to Edit and Update the VO if needed be. To save an edited VO, the information from the view interface is sent to the parser to convert it into proper format and the File Manager writes it to the XML Repository. The Delete Operation also works in the same fashion.

The Communication Manager acts as a server thread which listens for incoming connections from the remote client (SCM). Once is connection request is received, the

Communication Manager requests the VOM for VO information which is sent to the client.

3.2 Service Composition

Fig. 4 shows the static structure of the Service Composition Manager. It provides the overview of the main classes and the relationship, associations among these classes. The Form, TabControl and TabPage are the .Net built-in classes which act as displayable window and containers for visual controls respectively. The DeviceModule class provides the implementation of virtual representation for the input and output virtual objects. This is shown by the specialization relationship between the InputModule, OutputModule and the DeviceModule. The actual classes representing input devices such as a Pressure Sensor or an output device such as an LED are derived from the InputModule and OutputModule classes respectively. Each of these input or output device representation classes have associated custom attributes. The DeviceModule class implements the IDeviceModule interface for the implementation of core properties and methods related to devices modules. It also implements the ICloneable interface for making the virtual devices clone-able. This interface is used to clone the selected module when the user drags a module on the canvas.

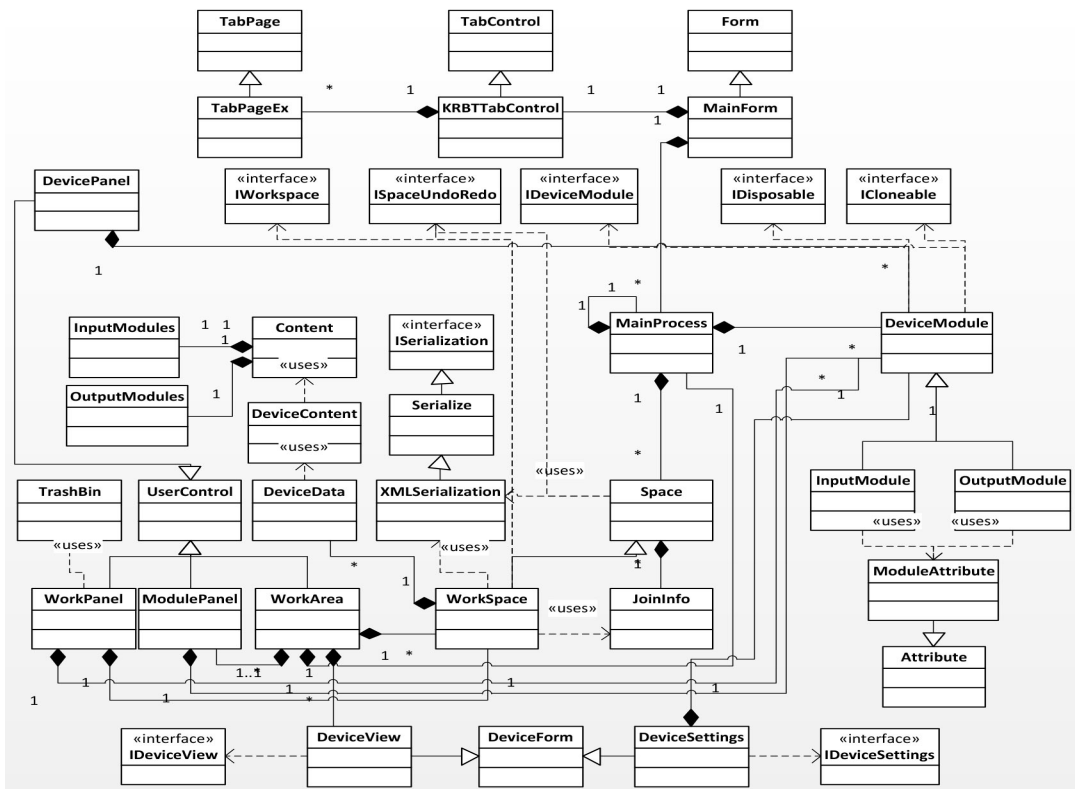


Figure 4: SCM static structure

Each device (VO) module such as the LED class has associated view and settings classes in the form of LEDView and LEDSettings classes. These classes are specializations from the DeviceView and DeviceSettings classes. The DeviceView class is associated with the WorkArea class to show the properties of a selected module in the form of Detail view tab in the editor. Similarly, the DeviceSettings class is a form for setting the properties or parameters and it is shown when a module is double-clicked in the editor's work area.

MainProcess class acts as the main back end process and it is implemented as a singleton class. All the other classes use the same instance of the MainProcess through a public interface. It maintains a list of DeviceModule class and Workspace class. Space class is super class of the Workspace class and it uses XmlSerialization class for the conversion of objects to XML data for storage purposes in the memory as well as the file system. The Space class implements the IWorkspace and ISpaceUndoRedo interfaces which contains the interfaces related to the storage of data space and maintenance of the current space by providing Undo and Redo functions respectively.

In order to maintain information about the joins created between the input and output modules drawn in the functionality editor, the Space class has a list of JoinInfo class. The same JoinInfo class list is used by the Workspace class to know the joins associated with the current project. Similarly, each Workspace object has an object of the DeviceData class which uses the DeviceContents and the Contents class for representing the input and output modules drawn on the work area of a given service composition project. The DevicePanel, WorkPanel and WorkArea classes are derived from UserControl class. These classes are used to provide the graphical user interface for individual projects which are displayed as objects of the TabPageEx class as tabs in the KRBTTabControl as an extended for of the TabControl class. TabPageEx is an extended version of the TabPage class which provides a close-able tabpage. The KRBTTabControl is part of the MainForm class which is the main displayable container for visual controls and components. The Trash class is a graphical representation of a waste bin which works with the WorkPanel class to provide the functionality for deleting a module drawn on the work area of the editor.

Fig. 5 shows the sequence of steps for designing or composing a service flow using the service composition manager. The user starts the main GUI first and then creates a new project. The first part of the figure shows the sequence of interaction among various internal components of the service composition manager when the user initiates a new project. This sequence does not show the initialization of the MainProcess. It is assumed that the editor is already initialized and the FrmMain container is already displayed on the screen where the user can click the new project button to start the process shown in this figure. As the user clicks the new project button on the FrmMain, it calls the createWorkArea() function and it in turn sends CreateSpace() message to the MainProcess. The MainProcess then creates an object of the Workspace class and returns it back to the FrmMain. The FrmMain then adds this new Workspace object to the spaces

collection of the MainProcess and further creates an object of the WorkArea class by passing the newly created WorkSpace object in the message. The WorkArea class has an associated WorkPanel object which actually acts as the drawing canvas for the SCM. It also creates the input and output panels for displaying the device module blocks that will be used by the user to drag-n-drop to the work area for creating the service design. To display this WorkArea object on the FrmMain, it is added to the controls collection of an extended tabpage object called as TabPageEx. This tabpage object is then displayed as a new tab on the tabcontrol and all the toolbar controls are setup for the new project using the enableControls message. At this point the user is displayed with the new project tab where he/she can drag and drop virtual objects to create the functionality flows. Once a new project is initialized, the user can start to “drag n drop” input and output VO onto the work area. As the work area has an associated WorkPanel control, the graphical representations for each VO dropped by the user is drawn on the panel.

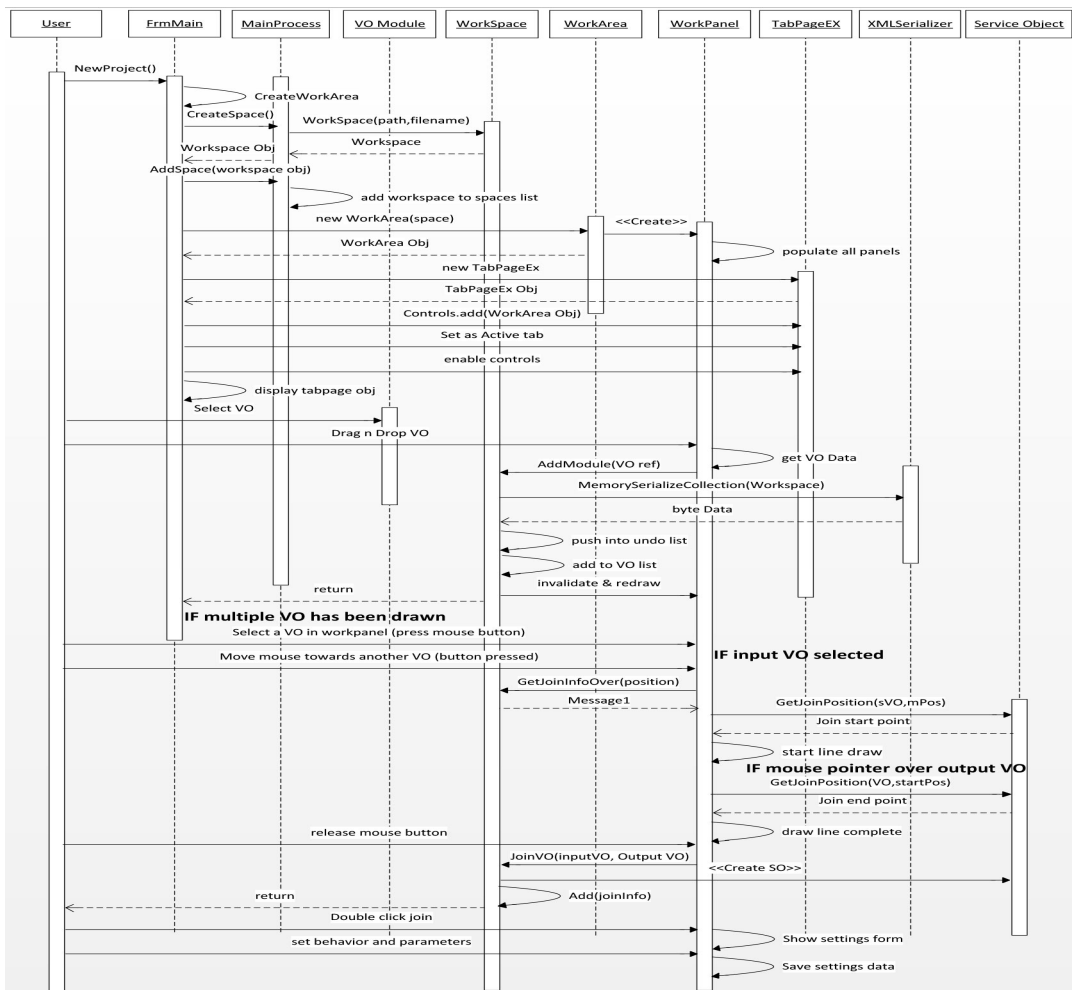


Figure 5: VO to SO mapping sequence at SCL

With each “drag n drop” on the panel, an event handler is executed which get the associated data of the dropped VO and creates a clone object from original one saved at the devices list at MainProcess. The clone object is then added to the Devices list maintained at each Workspace via the parent class Space. The parent class also has stack implementations for maintaining the Undo and Redo operations.

The Workspace class then creates an instance of the XmlSerialization class and calls the MemorySerializeCollection method with its own reference as parameter. The XmlSerialization class gets all the data associated with the Workspace object, converts it to XML format and saves it in a memory buffer as byte data. The reference to the byte data buffer is returned to the Workspace class. At this point, the byte data buffer is pushed into the Undo stack, the device list is updated and the WorkPanel is invalidated in order to draw the updated flow. This sequence is repeated for every new device module dropped onto the WorkPanel by the user. One thing is to be noted that each device module can be dragged to the WorkPanel only once in a project. This is an initial policy for the simplicity of the created flows and may be changed later on.

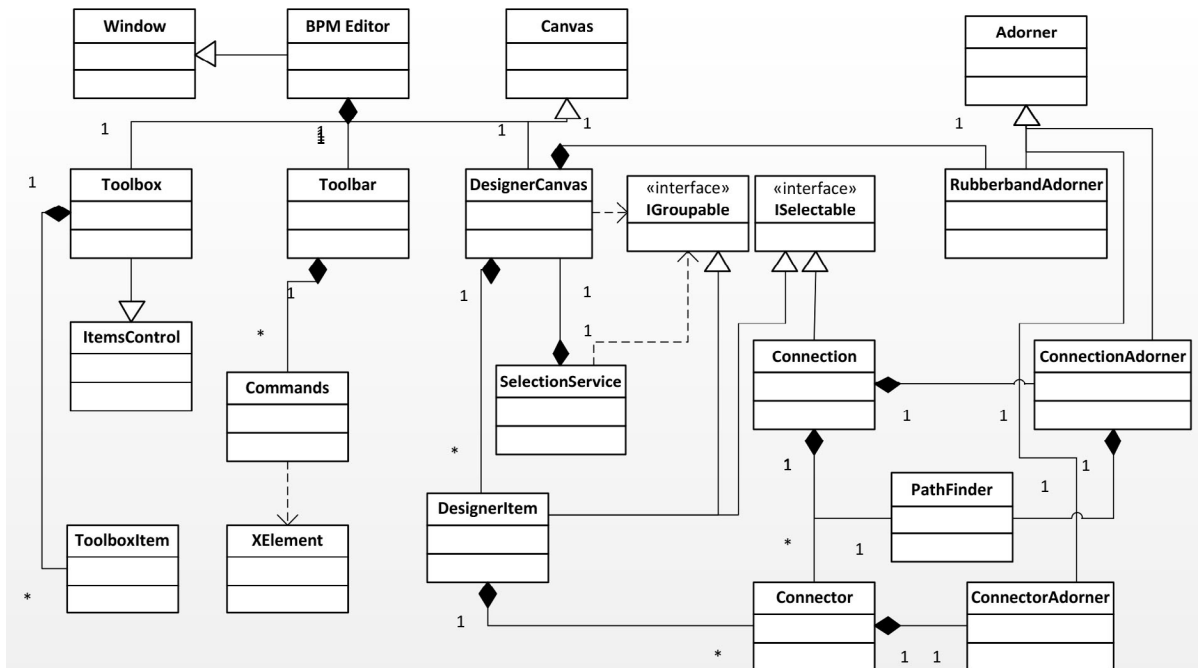


Figure 6: Static structure for business process design manager

The user can join an input device module to any number of output modules. As shown in the sequence, the user has to click and press the left mouse button on an input VO that is already drawn on the WorkPanel. If the user moves the mouse while the left button is pressed, the WorkPanel calls a static method of the JoinInfo class to get the starting

position from which the join should be drawn. To draw the join, a Bezier line is drawn from the starting point of the join to the mouse pointer. If the mouse pointer enters an output VO area, the static method is called again to calculate the end position of the join and the join is displayed on the WorkPanel. If the user releases the mouse button at this moment then the joinInfo object is created with the input and output device modules' information and the object is added to the Joins list maintained by the Workspace object. Otherwise, the drawn line is deleted. Once a join is created, the user can double click the join or the individual VO to set the behavior i.e. select an available function for the associated physical thing, and set other parameters. This data is saved as part of the SO in the form of JoinInfo and thus a complete SO is generated by combining input and output VOs.

3.3 Process Modeling

Fig. 6 shows the static structure of the main components at the business process layer. As the aim of the business process layer is to utilize the service objects created at the service composition layer and present them to the user in the form of business process modelling notations, the most important component at this layer is the business process design manager. It includes a BPM editor which is the main window containing the Toolbar, the Toolbox and the DesignerCanvas. The Toolbar is the component panel which is derived from the itemControl class. This panel is populated by the visual representations of the business process modeling notations in correspondence with the service objects acquired from the service composition layer.

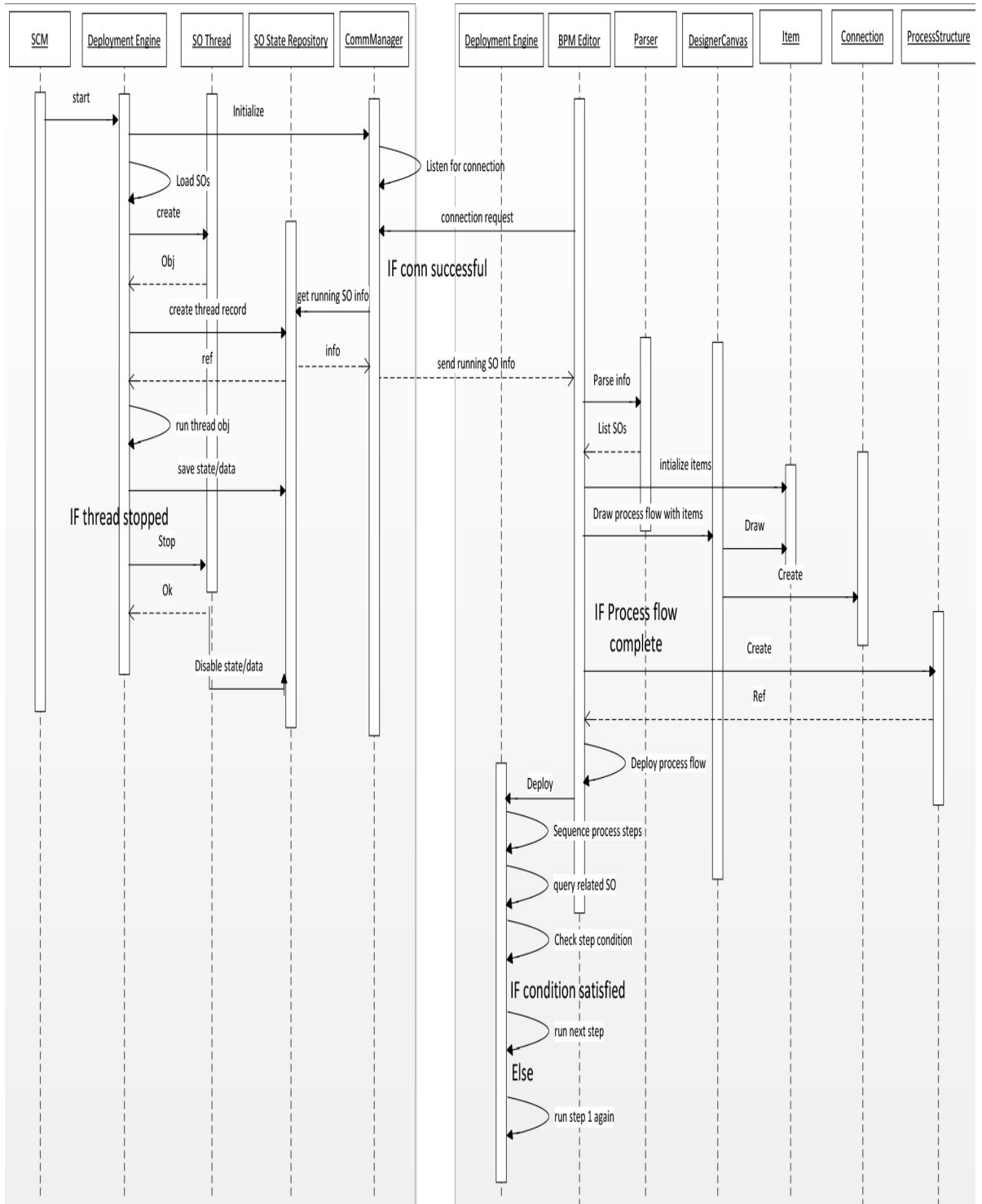


Figure 7: SO to process mapping and execution

The Toolbar component implements the basic drawing commands and operations which are required by a user for performing drag and drop designing. These commands include do, undo operations, copy, paste operation, group and ungroup operations etc. The connection class and its associated ConnectionAdorner class are used to draw connecting lines between the BPMN items representing the sequence of flow among process steps. Fig. 7 shows the sequence of operation at the business process layer. The operations at business process layer include the acquisition of SOs from the service composition layer, mapping them into business process modeling notations for the user to convert them into a process model and finally to execute the process. The figure shows that the SCM communication manager listens for connection requests from the BPM editor, the main component at the business process layer. Once the connection is successfully established, the available Service Objects (SOs) from the SO repository at the service composition layer are acquired in the form of XML info objects and sent to the business process layer. The acquired SOs are then parsed through an XML parser presented to the user as business process modeling notations. The user then creates the process model by visual drag and drop operations applied to the visual representations of business process model notations. The user draws the process components and connects them via the connection notation along with the operational rules or conditions applied for the transitions between steps. When the process model is complete, a process object is created. The process object is a deployable entity which is deployed via the Deployment Engine at the business process layer.

The deployed process object is converted into a sequence of operations based on the services which compose the service objects of the said process. The deployment engine interacts with the services as part of the process object. The data is acquired from the services, evaluated against the user set conditions/rules and the actuating services are executed as a result. The process object is run as a separate thread so multiple process objects can be executed simultaneously.

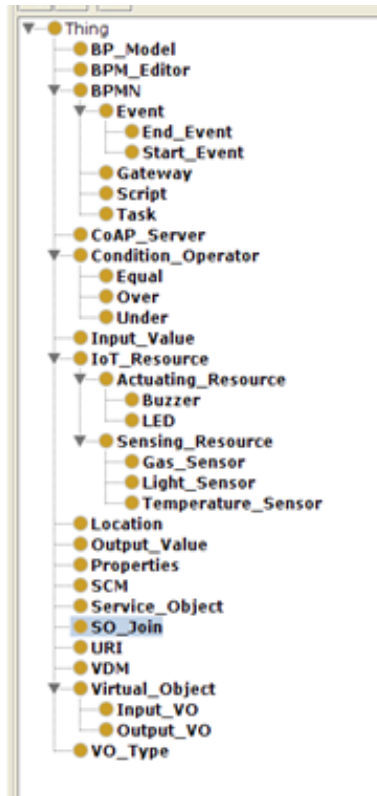


Figure 8: Semantic entities as part of the IoT application composition system

4. Semantic Modeling

This section provides the semantic modeling for the proposed system. The semantic models are used to allow a system to understand various concepts in the system’s operation and enable the system to infer extra knowledge by reasoning based on the provided information and rules. Given below is a protégé based implementation of the conceptual semantic entities and the relation among those entities.

4.1. Protégé Implementation

Thing is the base concept in the implementation of Internet of Things as well as in the field of semantics. Fig. 8 shows the core entities of the proposed system as part of the semantic model which is derived from the based class ‘Thing’. The semantic entities provide basic information in terms of specialization and generalization of entities. The relations are normally represented as ‘is-a’ relation.

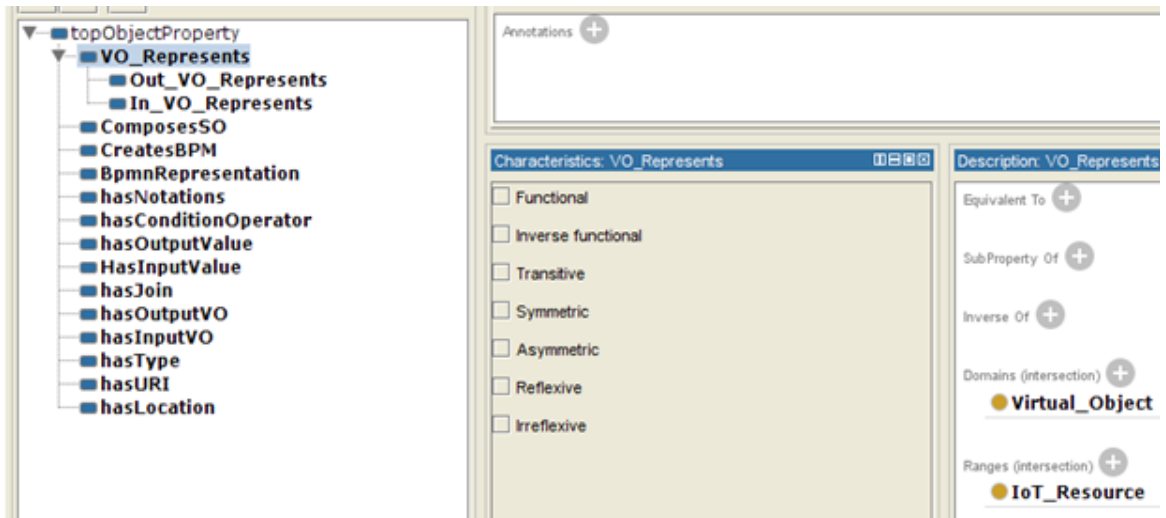


Figure 9: Object properties specifying semantic relations between

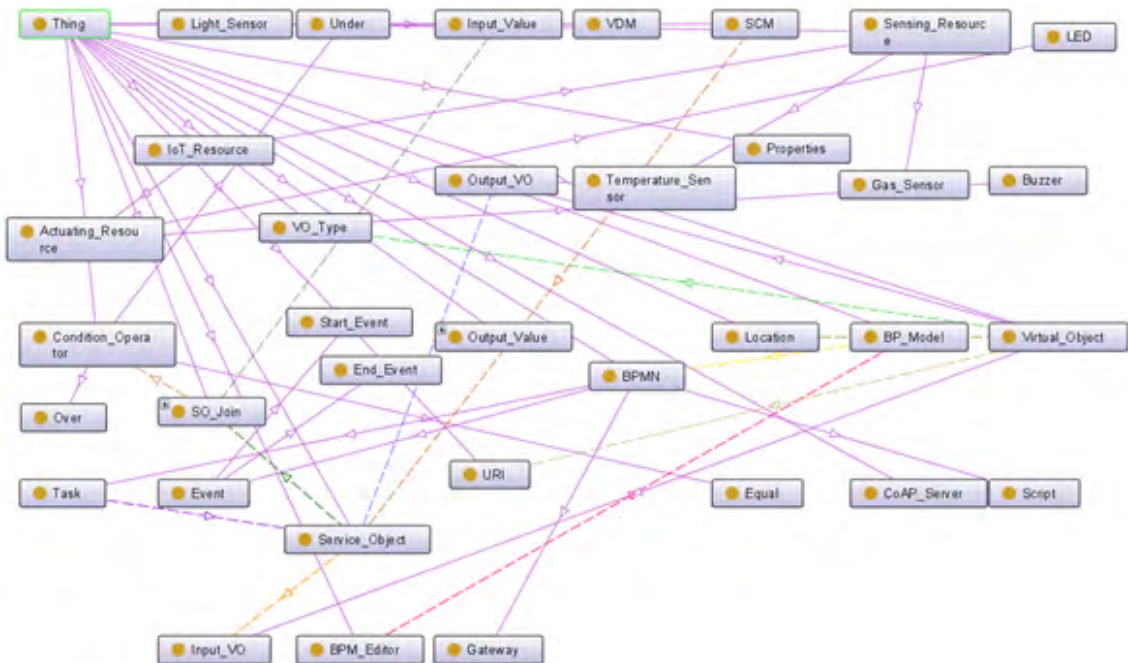


Figure 10: Protégé based semantic model for the proposed system

Fig. 9 provides a snapshot of the protégé object property tab. A list of the object properties defined based on the semantic entities has been shown on the left of the figure. These properties define relationships between various entities of the semantic model by specifying these entities as the domain and range of the properties. For example, the

VO_Represents object property specializes into two properties for the input and output VO concepts. The domain and range for generalized property has been specified to be the Virtual_Object and IoT_Resource entity respectively, providing the idea that a virtual object is a representation of an IoT resource. The specialization of the same object property then further divides this association between the input and output resources. The semantic model is the combination of semantic entities, relations among those entities and inference rules. Fig. 10 shows the semantic model for the proposed system as visualized by the protégé. Although the semantic model is still incomplete and further work is being done, it can provide semantic reasoning capabilities to the proposed system and based on the semantic inference, the system would be able to provide useful suggestions to the user while composing services as well as creating IoT applications through the BPM Editor.

4. Discussion:

The main focus of the architecture is the enablement of the end-users, with no or limited programming skills, to compose services/applications as per their own requirements with the help of the generic BPM notations. It is a layered architecture, which means that each layer can act on its own and only communicates with the upper or lower layers via well-defined interfaces. The layered architecture provides separation of concerns and avoid overburdening the constraint IoT devices. IoT devices exposes their functionalities as atomic services which are represented as virtual objects. These virtual object representations can then be grouped at the virtual object layer which makes the architecture scalable and the same can be achieved with the upper layers. The architecture can efficiently handle the security issues as well with the help of the separation of concerns. No user except the owner of the physical device is allowed to directly access and/or modify it. The users can only access virtual objects and service objects from different layers to fulfil their requirements and ultimately compose their own applications/services with the help of a generic drag and drop mechanism.

The only drawback of the layered architecture as evident from relevant literature is the performance degradation. When a message needs to pass through multiple layers in order to get across to the intended receiver, the extra translations and conversions at each layer degrades the overall performance. The future direction of the study is develop a prototype based on the presented architecture and test the performance at each layer and the performance of the composed services/applications.

5. Conclusion

This paper presented the idea of the utilization of business process modeling approach as a DIY tool for the IoT end users to design and execute IoT processes at IoT infrastructure. The IoT infrastructure may be in the form of a localized smart home or a more distributed implementation in the form of an agricultural IoT system. The paper visualizes the presented idea in the form of a layered architecture which consists of the physical layer,

service composition layer and the business process layer. A detailed description of the layered architecture and design detail of the layers has been presented in this paper. A brief description of the semantic model has also been presented. The development of the system is underway and a prototype implementation based on CoAP devices and services will be completed in the next phase of the development.

Acknowledgements

This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.10043907, Development of high performance IoT device and Open Platform with Intelligent Software). And this research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1017) supervised by the IITP (Institute for Information & communications Technology Promotion), Corresponding author; DoHyeun Kim (e-mail: kimdh@jejunu.ac.kr).

Author Biographies

Muhammad Sohail Khan received his Bachelor of Engineering and Master's degrees from Computer Software Engineering Department, University of Engineering and Technology Peshawar in 2008 and 2012 respectively. Meanwhile, he had been a part of the software development industry in Pakistan as a designer and developer. From 2010 onwards, he has been working as a faculty member at the Department of Computer Software Engineering, UET Peshawar, Pakistan. In August 2016 he received his Ph.D. degree from Jeju National University, South Korea. The focus of his work is the application of software design strategies towards enablement of mass involvement in IoT application/service development through intuitive DIY development environments.

Do Hyeun Kim received the B.S., M.S. and P.D degrees in Electronics Engineering from Kyungpook National University, Taegu, Korea, in 1988 and 1990, 2000 respectively. He joined the Agency of Defense Development (ADD), Korea, in 1990. Since 2004, he is currently a professor at the Department of Computer Engineering at Jeju National University, Korea. His research interests include sensor web, optimization algorithm and context prediction.

Faiza Tila received the B.Sc. degree in Computer Software Engineering from University of Engineering and Technology Pakistan in 2012. She joined the Mobile Computing lab Jeju National University as a M.S. student in 2014. Her area of interest is Semantic Web Technologies and Internet of things.

References

- [1] S. Sinha, "State of IoT 2021: Number of connected IoT devices growing 9% to 12.3 billion globally, cellular IoT now surpassing 2 billion" 2021. [Online] Available: <https://iot-analytics.com/number-connected-iot-devices/#:~:text=In%202021%20%2C%20IoT%20Analytics%20expects,than%2027%20billion%20IoT%20connections.> [Accessed: 25-April 2022].
- [2] V. Alpha, O. M. G. D. Number, and P. D. F. A. File, "BPMN 2.0 by Example," vol. 8, no. June, 2010.
- [3] N. Niknejad, I. Waidah, I. Ghani, B. Nazari, and M. Bahari, "Understanding Service-Oriented Architecture (SOA): A systematic literature review and directions for further investigation." *Information Systems* 91 (2020): 101491.
- [4] E. Schäffer, V. Stiehl, P. K. Schwab, A. Mayr, J. Lierhammer, and J. Franke. "Process-driven approach within the engineering domain by combining business process model and notation (BPMN) with process engines." *Procedia CIRP* 96 (2021): 207-212.
- [5] P. Harmon and C. Wolf, "State of Business Process Management – 2020," 2020.
- [6] M. R. Faheem, T. Anees, and M. Hussain. "The web of things: findability taxonomy and challenges." *IEEE Access* 7 (2019): 185028-185041.
- [7] S. Kumar, P. Tiwari, and M. Zymbler. "Internet of Things is a revolutionary approach for future technology enhancement: a review." *Journal of Big data* 6, no. 1 (2019): 1-21.
- [8] H. Garg, and M. Dave. "Securing iot devices and securely connecting the dots using rest api and middleware." In 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), pp. 1-6. IEEE, 2019.
- [9] M. Dave, J. Doshi, and H. Arolkar. "MQTT-CoAP Interconnector: IoT Interoperability Solution for Application Layer Protocols." In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 122-127. IEEE, 2020.
- [10] B. M. Presser, P. M. Barnaghi, U. Kingdom, M. Eurich, and C. Villalonga, "G l o b a l The SENSEI Project : Integrating the Physical World with the The IEEE ComSoc Sister Society in India : The Institution of Electronics and Telecommunication Engineers," no. April, pp. 1–4, 2009.
- [11] T. S. LOPEZ, D. KIM, G. H. Canepa and K. Koumadi, "Integrating Wireless Sensors and RFID Tags into Energy-Efficient and Dynamic Context Networks," vol. 52, no. 2, 2009.

- [12] S. Meyer, A. Ruppen, and L. Hilty, "The Things of the Internet of Things in BPMN," *Adv. Inf. Syst. Eng. Work.*, vol. 215, pp. 285–297, 2015.
- [13] M. Bauer, M. Boussard, N. Bui, and F. Carrez, "Project Deliverable D1.2 – Final Architectural Reference Model for IoT," no. 257521, pp. 53–59, 2013.
- [14] M. G., "Integrating the Internet of Things with business process management: A process-aware framework for Smart Objects," *CEUR Workshop Proc.*, vol. 1415, pp. 56–64, 2015.
- [15] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.
- [16] K. Gama, L. Touseau, and D. Donsez, "Combining heterogeneous service technologies for building an Internet of Things middleware," *Comput. Commun.*, vol. 35, no. 4, pp. 405–417, Feb. 2012.
- [17] L. Atzori, A. Iera, and G. Morabito, "From 'smart objects' to 'social objects': The next evolutionary step of the internet of things," *IEEE Commun. Mag.*, vol. 52, no. 1, pp. 97–105, Jan. 2014.
- [18] C. Anderson, "Makers: The New Industrial Revolution," *Compet. Rev.*, vol. 24, no. 2, pp. 147–149, Mar. 2014.
- [19] J. G. Tanenbaum, A. M. Williams, A. Desjardins, and K. Tanenbaum, "Democratizing technology," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 2013, p. 2603.
- [20] D. Carrera, "Collaborative Open Market to Place Objects at your Service," pp. 1–65, 2014.
- [21] IDC, "Worldwide Internet of Things Spending Guide". 2022 [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=IDC_P29475. [Accessed: 25-April 2022].

The Impact of Mitigation Strategies on Geographical Distance Issues in GSD: An Empirical Evaluation

Nadia Ka e nat¹

Uzair Iqbal Janjua²
Atta ur Rahman⁴

Tahir Mustafa Madni³

Abstract

Geographical distance is one of the most critical communication challenges in global software development projects; that significantly affects the projects' quality, cost, and schedule. Thus, it leads the project toward failure. Mitigation practices may help organizations overcome geographical distance challenges. In the past, the authors of this study conducted a systematic literature review to identify the geographical distance challenges and their relevant mitigation strategies to propose a conceptual framework. It is difficult to explain the exact relationship between geographical distance challenges and mitigation strategies without empirical analysis. Therefore, the main objective of this study is to empirically evaluate the proposed conceptual framework and to analyze the impact of identified mitigation strategies on geographical distance risks. The finding of this study shows that different mitigation strategies have a significant impact on different geographical distance risks with a $p\text{-value} < 0.01$. Based on the results, this research can help software organizations to tackle geographical distance challenges by using appropriate mitigation strategies to reduce the software project's failure rate.

Keyword: Communication Challenges, Global Software Development, Distributed Software Development, Empirical Evaluation, Geographical Distance Risks, Mitigation Strategies.

1. Introduction

Over the last decade, many software organizations around the globe have started adopting global software development (GSD) due to its profound benefits such as low development cost and time, access to cheap skilled labor, etc. [1], [2]. GSD is carried out by knowledge team members in different geographical locations worldwide to develop commercially competitive software for organizations [3]. While working globally, GSD proved beneficial for software organizations [4]. Many developing countries including India, Afghanistan, Thailand, and Pakistan contribute to GSD activities to create a product

¹COMSATS University Islamabad, Islamabad Pakistan | nk_rehman@yahoo.com

²COMSATS University Islamabad, Islamabad Pakistan | uzair_iqbal@comsats.edu.pk

³COMSATS University Islamabad, Islamabad Pakistan | tahir_mustafa@comsats.edu.pk

⁴COMSATS University Islamabad, Islamabad Pakistan | attaurrahman513@gmail.com

for global market within quality, budget, and schedule constraint [5]. As a result, GSD has become an emerging paradigm for developing the software system in the IT industry. Despite these benefits, organizations are facing several challenges in GSD as well. The Communication challenge, the most critical one [1], is the primary cause of software project failure [2]. There are three type of communication challenges that threaten the incentives of the GSD. These challenges are Geographical distance, Temporal distance, and Cultural distance [1]. Temporal distance is the time zone difference between teams working at the distributed location. Cultural distance is the understanding of language, religion, and organizational culture of other team members working at remote locations, and geographical distance is defined as “effort required for one team member to visit another.” Geographical distance risks occur because of the dispersion of team members over several distant locations [1],[6]. Among these communication risks, geographical distance is the most significant risk [6],[7],[8]. It has a ripple effect on other challenges such as cultural and temporal distance [9]. It causes delays and misunderstandings among distributed team members [10] because the amount of the information provided to dispersed team members is limited. Therefore, it is necessary to look for mitigation strategies to reduce the potential impact of these risks [11]. A few researchers also proposed different mitigation strategies to address geographical distance challenges [11],[12],[13].

Some of the effective communication strategies that will help reduce the negative impact of geographical distance issues are traveling between sites [14], Promoting informal communication [11], and Synchronous communication is, the most probable solution to alleviate the negative impact of communication risk. As team members have limited opportunities to meet face to face, usage of synchronous devices helps reduce misunderstandings between dispersed teams.

The author in [1] discusses communication issues and proposes a conceptual framework, but mitigation strategies are not discussed, and the framework is not empirically validated. In the study [8], the author gives strategies to address geographical distance challenges, but these guidelines are not empirically validated. The author in [15] identifies challenges posed by geographical distance and their relevant mitigation strategy through SLR and proposes a conceptual framework. However, to the best of our knowledge, no empirical study has been performed to evaluate the framework and analyze the impact of mitigation strategies on geographical distance issues. Therefore, the lack of empirical investigation leads to a gap in the existing literature. To fill the gap, the authors of this study proposed a conceptual framework in the past, and in this study, the proposed conceptual framework is empirically validated by small and medium-sized (SMEs) GSD organizations of Pakistan. The remaining section of this paper is organized as follows: In section 2 literature review is discussed, and section 3 discusses the research methodology, Section 4 discusses the result, and discussion of the current study is provided in section 5, Finally, in section 6 the

conclusion and future work are discussed.

2. Literature Review

In the study [12], authors conducted an SLR to identify communication risks and discuss general solutions to overcome these challenges. In [7], authors surveyed to determine the effect of geographic distance on software development organizations. The survey results show that geographic distribution damages information sharing and communication channels among distributed software organizations. Another study conducted an SLR to identify risks for communication and provide solutions to overcome those challenges. However, the author did not propose any framework, and empirical evaluation was not performed[9]. Moreover, in [13], Communication risks and mitigation strategies during requirement change management in GSD are identified. The framework is also proposed, but the framework is not empirically validated. Furthermore, the author in [11] prioritized geographical issues and mitigation strategies with the help of the ANP algorithm. The author in [16], conducted an SLR to find out communication risk and mitigation strategies for the requirement engineering process in GSD. An author in [15], conducted a systematic literature review and identify eight issues that are caused by geographical distance and their relevant mitigation strategies. After that a conceptual framework is proposed. However, the framework is not empirically validated.

According to results of SLR, conducted in [15], most of the studies discuss that geographical distance risk is the most significant communication risk as compared to other risks [6], [7], and [8]. In a nutshell, we did not find any study that empirically evaluated the impact of mitigation strategies on geographical distance challenges in the GSD context. Table 1 shows a summary of the literature review.

Table 1: Summary of Literature Review

Reference	Description	Methodology	Evaluation	Limitation
[9]	Communication risks and Mitigation strategies are discussed.	SLR	NOT	Empirical evaluation is required.
[13]	Framework is proposed for communication risks during requirement change management process and their relevant mitigation strategies.	SLR	NOT	Empirical evaluation is required.
[16]	Communication risk and their mitigation strategies are discussed.	SLR	NOT	Empirical evaluation is missing.
[15]	Framework is proposed for geographical distance risks and their relevant mitigation strategies.	SLR	NOT	Empirical evaluation is missing.

3. Research Methodology

This section explains the research methodology of the current study. The Overall research

design is shown in Figure 1. Given below are the steps used to conduct the study.

A. *Systematic Literature Review*

In our previous study, SLR was conducted to extract geographical distance issues for communication and their relevant mitigation strategies from literature. A total of eight geographical distance issues were identified and nine common strategies were extracted for their respective risks. These mitigation strategies resolve more than one issue. After completing SLR a conceptual framework was proposed [15]. Figure 2 shows the proposed conceptual framework.

The proposed conceptual framework is a second-order formative model. To evaluate the proposed conceptual framework, suggested formative measures were applied. There are two-second order formative constructs, i.e., Mitigation practices for geographical distance issues in GSD and Geographical Distance issues. Both second-order constructs are further composed of 8 different first-order constructs. Each first-order construct has unique properties different from others, so removing any item is omitting a part of the construct.

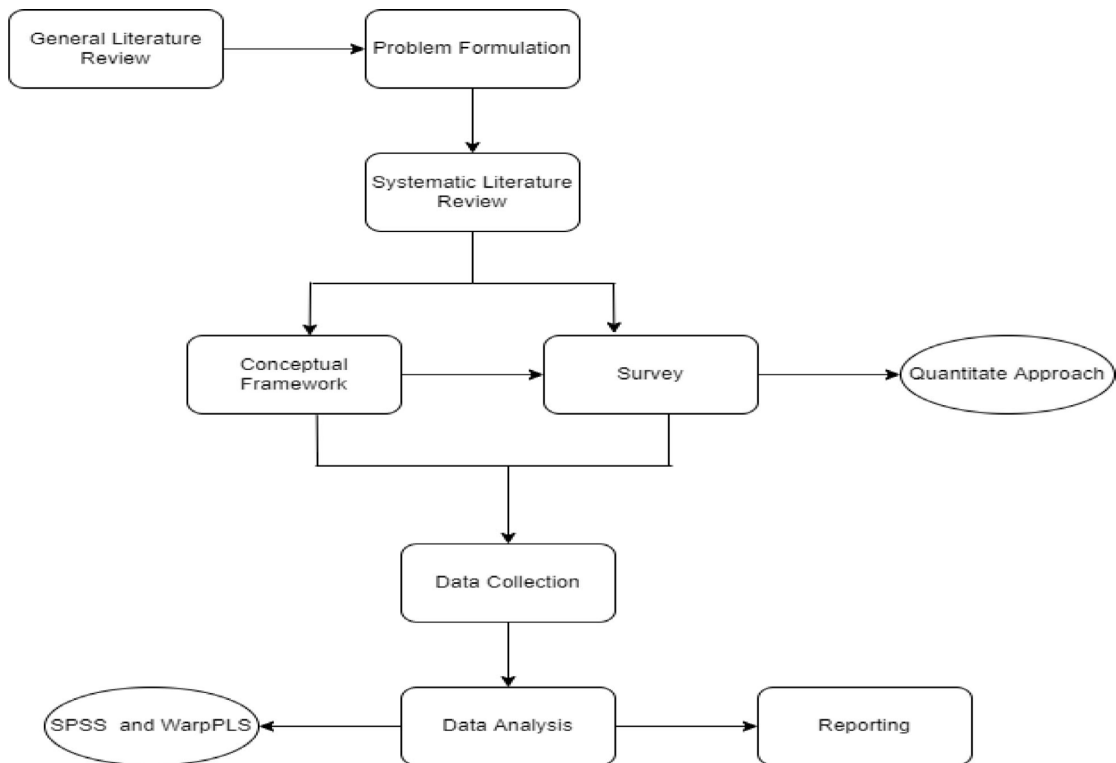


Figure 1: Research Methodology

B. Hypothesis Development

Following the proposed conceptual framework, a total of 9 hypotheses are hypothesized and given below.

H1: Lack of trust mitigation strategies (MSLOT) positively impacts geographical distance issues in GSD.

H2: Lack of team cohesiveness mitigation strategies (MSLOC) positively impacts geographical distance issues in GSD.

H3: Lack of informal communication mitigation strategies (MSLFFM) positively impacts geographical distance issues in GSD.

H4: Lack of interpersonal relationship mitigation strategies (MSLIC) positively impacts geographical distance issues in GSD.

H5: Loss of communication richness mitigation strategies (MSLIR) positively impacts geographical distance issues in GSD.

H6: Communication frequency reduced mitigation strategies (MSLCR) positively impacting geographical distance issues in GSD.

H7: A Communication effort increase mitigation strategy (MSCEI) positively impacting geographical distance issues in GSD.

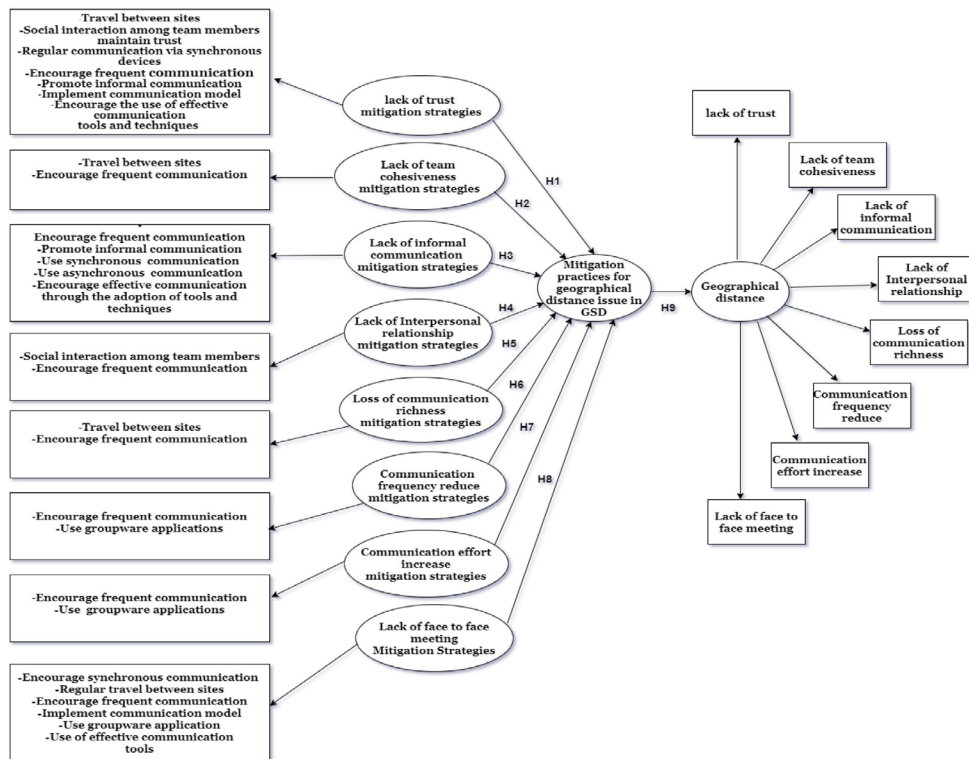


Figure 2: Conceptual Framework for Geographical Distance Issues and their Mitigation Strategies in GSD

H8: Lack of face-to-face meeting mitigation strategies (MSLCFR) positively impacts geographical distance issues in GSD.

H9: Mitigation practices have a positive impact on geographical distance issues.

C. *Empirical Analysis of Conceptual Framework*

This section, presents the empirical analysis of the conceptual framework.

Measure and Procedure for Data Collection

A quantitative research method was used in this study to investigate the geographical distance issues in GSD. A closed-ended questionnaire was developed and used to obtain GSD-based organization's data to evaluate and test the conceptual framework. The questionnaire consists of 3 main sections shown in Figure 3. The ordinal scale is used to understand the relative rank of variables. The option include in scale are starting from 0 = "No contribution at all", 1 = "slightly contributive", 2 = "Moderately contributive",

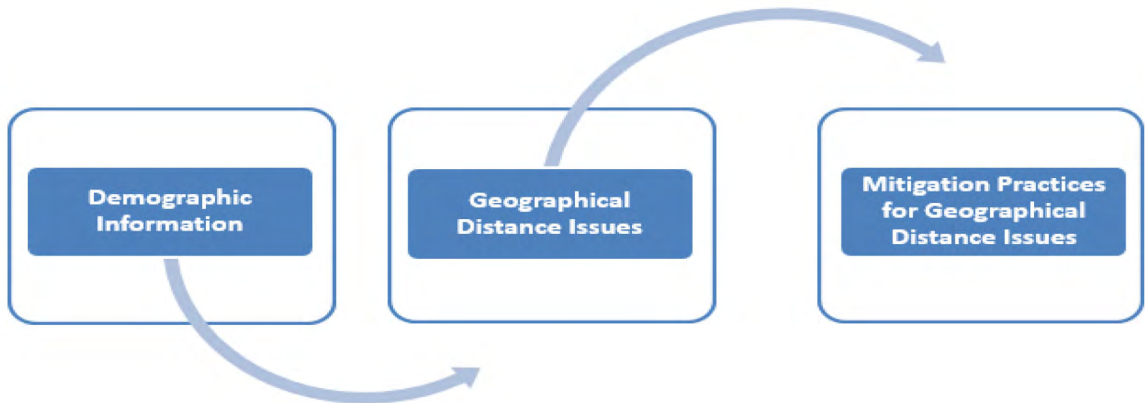


Figure 3: Survey Design

1). *Participants*

After pilot study the convenience sampling technique was used in this research study because all organizations in Pakistan are not GSD-based. Only GSD-based organizations have been targeted for data collection. Data was collected from April 11 to May 7, 2021. The survey was distributed online using the LinkedIn platform to 400 people, out of which 212 people replied and filled the survey. A total of six questionnaire were discarded as they were not filled correctly. A total of 206 responses received, yielding a 51 percent of response rate in final study.

2). *Data Analytical Approach*

After the Cronbach's alpha test, correlation analysis, we used the PLS-SEM method to test the hypotheses as it helps to analyze second-order formative construct. The PLS-SEM method consists of two sub-model, i.e., the structural and measurement model. The structural models show the relationship between dependent and independent variables. On the other hand, the measurement model depicts the relationship between variables and data collected with the help of a survey [17].

4. **Result of an Empirical Analysis**

This section presents the finding of the empirical investigation. We examined each hypothesis and also analyzed its outcome.

1). *Demographic Profile of Respondent*

The suggested sample for PLS-SEM use is 200 or above [18]. Therefore, a total of 206 responses were collected in this research study. Table 2 lists respondent's demographic information.

Table 2: Summary of Respondent demographic information

Demographics	Respondent	Frequency	Percentage
Gender	Male	198	96.1%
	Female	8	3.9%
Total	-	206	100%
Education	Diploma	0	0
	Bachelor's	152	73.8%
	Master's	54	26.2%
Total	-	206	100%
Work experience	1-4 years	134	65%
	5-9 years	48	23.3%
	More than 10 years	24	11.7%
Total	-	206	100%
Role	Developer	130	63.1%
	Analyst	3	16%
	Tester	33	5.9%
	Test Manger	12	10.2%
	Project Manager	21	0
	Designer	0	0.4%
	EO	1	2.9%
	Others	6	1.5%
Total	-	206	100%

No of employees	Between 10-25	13	6.3%
	Between 26-50	6	2.9%
	Between 51-100	21	10.2%
	Between 100-250	166	80.6%
Total	-	206	100%

1). *Reliability Analysis of Questionnaire*

Cronbach alpha test was applied to test the reliability of the questionnaire i.e. to check internal consistency among the variable of the questionnaire. According to [18], the minimum value of 0.6 is acceptable. Table 3 shows the results of the Cronbach alpha test.

Table 3: Cronbach Alpha Result

Construct	Items	Cronbach Alpha
MSLOT	7	0.875
MSLOC	2	0.620
MSLIC	5	0.862
MSLIR	2	0.668
MSLCR	2	0.660
MSCEI	2	0.673
MSCFR	2	0.68
MSLFFM	6	0.881

2). *Correlation Analysis*

In this section, correlation analysis among the construct was analyzed and discussed. Correlation analysis was conducted between dependent and independent variables using SPSS before performing PLS-SEM analysis. According to [19], for correlation analysis coefficient value must lie between +1 and -1. If the value of correlation is greater than 0.8, then a strong correlation exists between variables [20]. Table 4 summarizes the proposed conceptual framework's correlation analysis among independent and dependent variables. According to the results, a strong correlation exists between variables. Because of that, the estimated level of collinearity is very high. Collinearity is a serious threat in a formative model. In the formative model, there should be no high intercorrelation between variables. High collinearity between variables affects the significance of overall results [21]. According to [22], the acceptable value of collinearity should be less than 3.3. After analyzing our model by using WarpPLS version 6.0, the collinearity between variables is 7.041. Because of that overall significance of the result has been compromised. To test the framework and to check the impact of independent variables on dependent variables, we split a conceptual framework into eight models and checked the significance of each

mitigation strategy against their relevant geographical distance issue. These Models are shown in Figures starting from 4, to 11.

Table 4: Correlation Analysis

	GDI	MS LOT	MS LOC	MS LFFM	MS LIC	MS LIR	MS LCR	MS CEI	MS CFR
GDI	1								
MSLOT	0.901**	1							
MSLOC	0.803**	0.780**	1						
MSLFFM	0.915**	0.844**	0.781**	1					
MSLIC	0.906	0.785**	0.687**	0.855**	1				
MSLIR	0.839	0.781**	0.704**	0.782**	0.770**	1			
MSLCR	0.815	0.724**	0.739**	0.740**	0.705**	0.741**	1		
MSCEI	0.847	0.758**	0.631**	0.771**	0.777	0.743**	0.717**	1	
MSCFR	0.615	0.455**	0.396**	0.475**	0.517	0.494**	0.438**	0.568**	1

** Correlation significant at 0.01 level

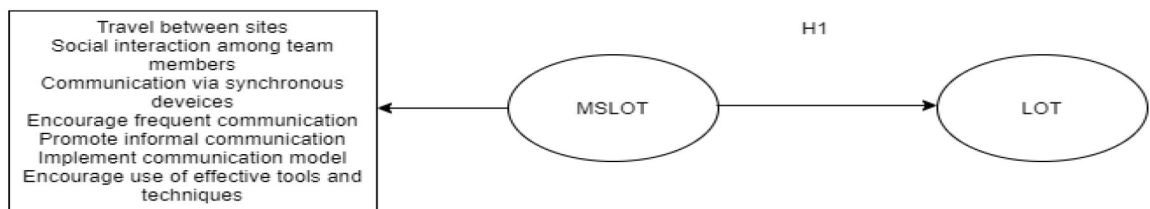


Figure 4: Model for MSLOT

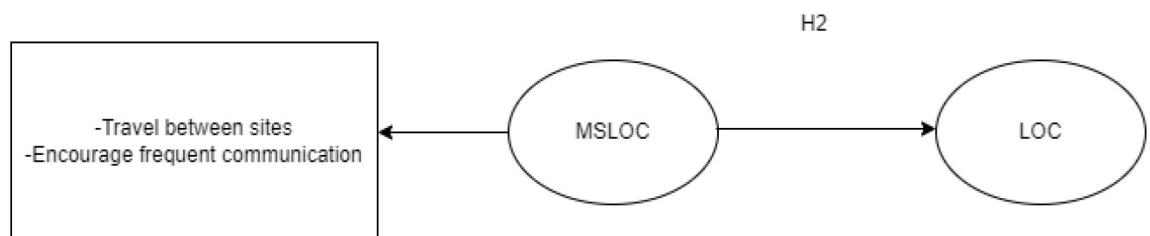


Figure 5: Model for MSLOC

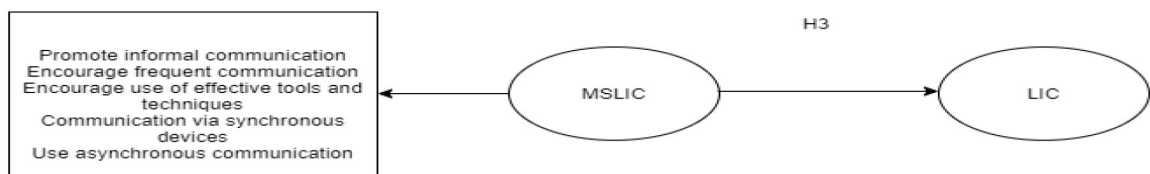


Figure 6: Model for MSLIC

H4

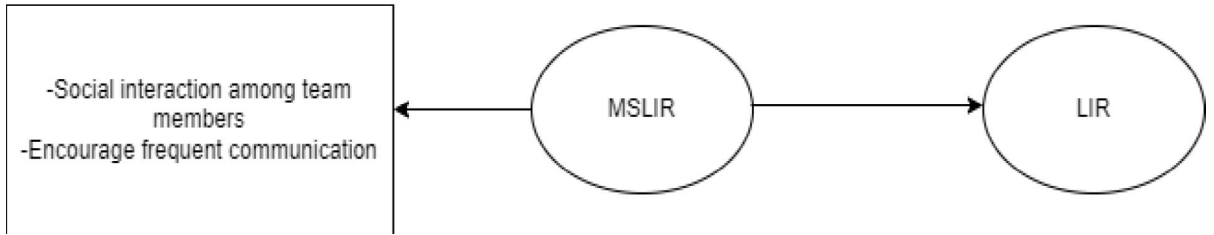


Figure 7: Model for MSLIR

H5

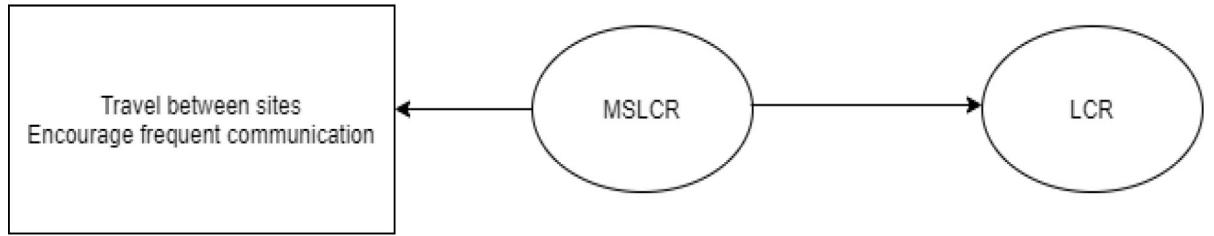


Figure 8: Model for MSLCR

H6

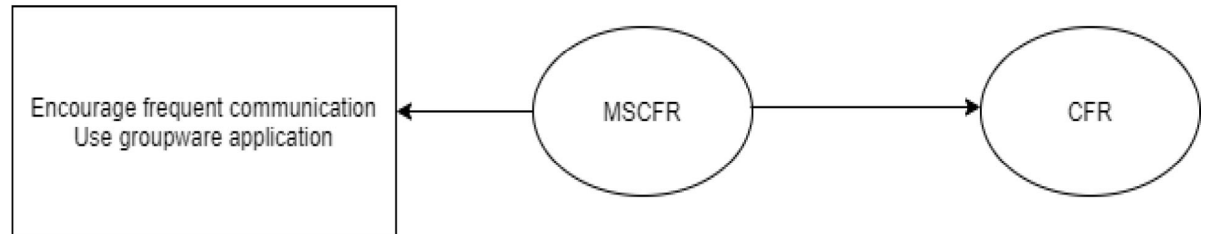


Figure 9: Model for MSCFR

H7

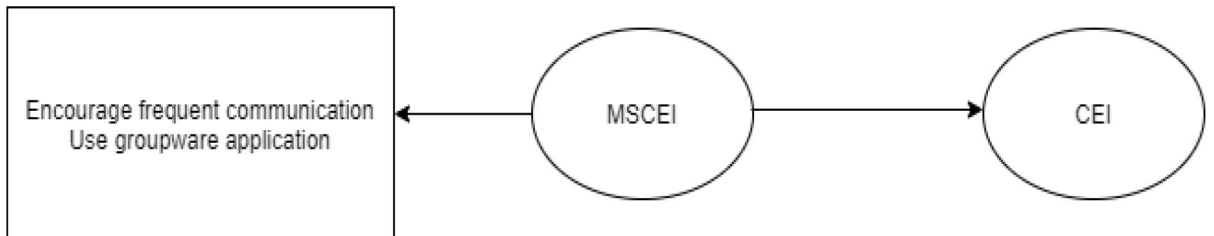


Figure 10: Model for MSCEI

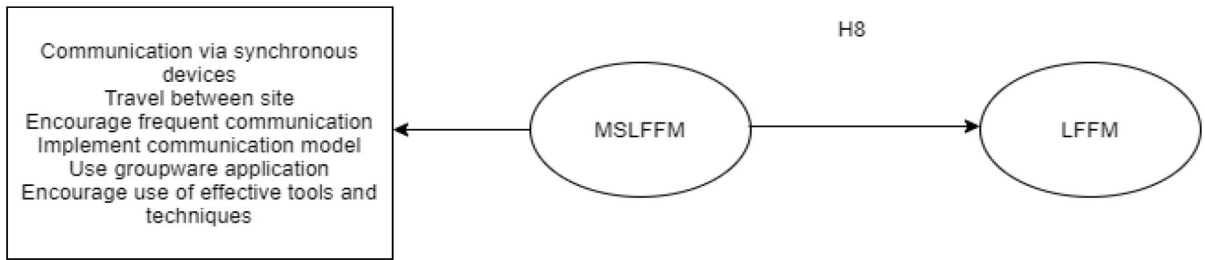


Figure 11: Model for MSLFFM

4). Model Assessment

PLS-SEM was applied in this study. Initially, to describe the accuracy and authenticity of the construct measurement model was accessed. Then, a structural model was accessed that describes the significance of the relationship or association between the constructs.

A. Assessment of Measurement Model

PLS Mode B algorithm is more suggested for formative measurement assessment [22]. Therefore, the PLS Model B algorithm was used in this study. Firstly, the variance inflation factor (VIF) is used to evaluate the construct's validity. After that R-square, beta coefficient, loading, weight and P-value was acquired.

- VIF is acceptable if its value is less than 5 and ideal if, its value is less than 3.3 [21].
- Loading, weight, VIF, full collinearity, and significant level of items was accessed to check the reliability of the formative construct.
- All items were acceptable if their loading value is greater than 0.5 [21].
- "R-Square represent the percentage of variance in independent variable caused by dependent variable "[23].
- Beta value compare the strength of each individual variable on dependent variable. The higher the value of beta coefficient the higher is the effect [23].
- P-value shows relationship significance if its value less than 0.05. We can say that relationship among variables is significant [23].

Table 5 shows the result for measurement model assessment. Items(column) represent list of all mitigation's strategies against each issue. Evaluation of the measurement model indicates that all constructs are statistically significant.

Table 5: Evaluation of Formative Measurement Model

Construct	Items	Loading	Weight	Significance	Full-Collinearity	Beta	R-square	VIF		
MSLOT	M1LOT	0.491	0.042	<0.01	1.380	0.53	0.28	1.328		
	M2LOT	0.832	0.362	<0.01				1.921		
	M3LOT	0.862	0.347	<0.01				2.471		
	M4LOT	0.800	0.172	<0.01				3.112		
	M5LOT	0.571	0.241	<0.01				2.306		
	M6LOT	0.795	0.136	<0.01				3.121		
	M7LOT	0.852	0.318	<0.01				2.961		
MSLOC	M1LOC	0.961	0.801	<0.01	1.625	0.63	0.39	1.330		
	M2LOC	0.719	0.321	<0.01				1.330		
MSLFFM	M1LFM	0.803	0.281	<0.01	1.817	0.57	0.57	1.817		
	M2LFM	0.669	0.245						<0.01	1.344
	M3LFM	0.869	0.286						<0.01	2.945
	M4LFM	0.820	0.188						<0.01	2.500
	M5LFM	0.774	0.240						<0.01	2.046
	M6LFFM	0.781	0.029						<0.01	2.887
	M7LFFM	0.781	0.029						<0.01	2.887
MSLIC	M1LIC	0.790	0.274	<0.01	1.858	0.68	0.46	2.175		
	M2LIC	0.682	0.201	<0.01				2.514		
	M3LIC	0.857	0.373	<0.01				2.389		
	M4LIC	0.849	0.414	<0.01				2.019		
	M5LIC	0.773	0.322	<0.01				1.699		
MSLIR	M1LIR	0.897	0.581	<0.01	1.875	0.70	0.48	1.506		
	M2LIR	0.881	0.544	<0.01				1.506		
MSLCR	M1LCR	0.741	0.428	<0.01	2.062	0.72	0.52	1.217		
	M2LCR	0.922	0.741	<0.01				1.217		
MSCEI	M1CEI	0.866	0.529	<0.01	1.713	0.65	0.42	1.456		
	M2CEI	0.899	0.603	<0.01				1.456		
MSCFR	M1CFR	0.937	0.861	<0.01	1.515	0.59	0.34	1.049		
	M2CFR	0.542	0.356	<0.01				1.049		

B. Assessment of Structural Model

To evaluate the structural model, hypotheses of the proposed conceptual framework were tested, and the significance of the construct was evaluated using WarpPLS version 6.0. The acceptable p-value is <0.05. Table 6 shows the assessment of the structural model. The table shows that lack of trust (LOT) mitigation practices has a significant

impact on geographical distance issues with a p-value less than 0.01. Moreover, lack of team cohesiveness (LOC) mitigation practices significantly impacts geographical distance issues with a p-value less than 0.01. The Lack of face-to-face meeting (LFFM) mitigation practices significantly impacts geographical distance issues with a p-value less than 0.01. Similarly, Lack of informal communication (LIC) mitigation practices significantly impacts geographical distance issues with a p-value less than 0.01. Lack of interpersonal relationship (LIR) mitigation practices significantly impacts geographical distance issues with a p-value less than 0.01. Also, Loss of communication richness (LCR) mitigation practices significantly impacts geographical distance issues with a p-value less than 0.01. Moreover, Communication effort increase (CEI) mitigation practices significantly impact geographical distance issues with a p-value less than 0.01. Communication frequency reduced (CFR) mitigation practices significantly impact geographical distance issues with a p-value less than 0.01. Overall, mitigation practices significantly impact geographical distance issues with a p-value less than 0.01. The hypotheses proposed in the Hypothesis development section are supported and approved based on the results.

The more the organization uses these mitigation strategies in their GSD projects, the negative impact of issues will be reduced.

Table 6: Evaluation of Formative Structural Model

Hypothesis Testing	P-value	Results
H1: MSLOT → LOT	<0.01	Supported
H2: MSLOC → LOC	<0.01	Supported
H3: MSLIC → LIC	<0.01	Supported
H4: MSLIR → LIR	<0.01	Supported
H5: MSLCR → LCR	<0.01	Supported
H6: MSCFR → CFR	<0.01	Supported
H7: MSCEI → CEI	<0.01	Supported
H8: MSLFFM → LFFM	<0.01	Supported
H9: MS → GDI	<0.01	Supported

5. Discussion

In the current study, it has been observed that communication between distributed team members is hampered because of geographical distance risks. Geographical distribution influences the essence of team interactions and provides fewer opportunities for spontaneous interaction and team knowledge acquisition. Communication among dispersed team members becomes more complex as geographical distance increases. In our previous study [15], a conceptual framework was proposed, which was formative

and empirically evaluated in this study. For empirical evaluation, a survey was conducted in which more than 200 participants from GSD SMEs participated.

PLS-SEM is used to perform statistical analysis. According to the result of correlation analysis shown in table 4, a strong correlation exists between variables ($r > 0.85$); because of that, the estimated level of collinearity is very high. Collinearity is a severe threat in a formative model, as in formative model, there should be no intercorrelation between variables. High collinearity between variables affects the significance of overall results. After analyzing our model using WarpPLS version 6.0, the collinearity between variables is 7.041. Because of that overall significance of the result has been compromised. To test the framework and check the impact of independent variables on dependent variables, we split it into eight models and check the significance of each mitigation strategy against their relevant geographical distance issue. Nine hypotheses were developed to examine the impact of independent variables on dependent variables. Hypothesis testing was done with the help of PLS-SEM. Initially to check authenticity and accuracy of each construct measurement model was accessed. To describe the significance of relationship between dependent and independent variable structural model was accessed. There result is shown in, table 5 and table 6. Each hypothesis is addressed and discussed separately, which is given below.

- **H1**

LOT is the leading risk that affects communication in GSD. Mitigation practices help to reduce the potential effect of these issues. The beta value for MSLOT is obtained as 0.53, R square is 0.28 with items loading greater than 0.5. The P-value of the overall construct is less than 0.01, which is statistically significant. This implies that MSLOT helps to minimize LOT issues in the GSD environment. Therefore, H1 is supported in this research.

- **H2**

The relationship between mitigation strategies and LOC issues could be shown by beta value, and R-square value obtains as 0.63, 0.39 with items loading greater than 0.5. The P-value for construct is < 0.01 (< 0.05), showing the significant impact of MSLOC on LOC geographical distance risk. Therefore, H2 is supported in this research based on the above relationship result.

- **H3**

The beta value and R-square value of MSLIC are obtained as 0.68 and 0.46, with items loading > 0.5 . The obtained p-value < 0.01 ($p < 0.05$), which shows a positive impact of MSLIC on the LIC issue. Therefore, H4 is supported in this research.

- **H4**

The relationship between mitigation strategies and LIR issues could be shown by beta value and R-square value obtained as 0.70 and 0.48 with items loading greater than 0.5. P-value is obtained as < 0.01 (< 0.05), which shows significance. Hence, H6 is supported in this research.

• H5

The relationship between mitigation strategies and LCR issues could be shown by beta value and R-square value obtained as 0.72 and 0.52 with items loading greater than 0.5. P-value is obtained as <0.01 (<0.05), showing the significant impact of MSLCR on LCR. Therefore, result H5 is supported in this research based on the above relationship.

• H6

The relationship between mitigation strategies and CFR issues could be shown by beta value and R-square value obtained as 0.59 and 0.34 with items loading greater than 0.5. P-value is obtained as <0.01 (<0.05), showing the significant impact of communication frequency reduced mitigation strategies on CFR geographical distance risk. Therefore, H8 is supported in this research.

• H7

The beta value and R-square value of MSCEI were obtained as 0.65 and 0.42, which show a positive impact of mitigation strategy on a dependent variable with items loading greater than 0.5 with a p-value <0.01 that show a significant relationship. Therefore, H7 is supported in this research.

• H8

According to the result, the beta and R-square values were obtained as 0.75 and 0.57, showing a positive impact of MSLFFM on LFFM risk. As LFFM is one of the important risks that cause GSD communication issues, it is necessary for the organization to choose a suitable mitigation strategy to cope with this issue. The loading value of all MSLFFM items is greater than 0.5. The significant impact of lack of face-to-face mitigation strategy on the LFFM issue was shown by p-value <0.01 . Hence, H3 is supported.

• H9

To check the impact of mitigation strategies on geographical distance issues, hypothesis H9 was tested. The result shows that p-value is less than 0.01, which shows the significant impact of mitigation strategies on geographical distance issues. Hence, based on the above relationship, result H9 is supported in this research.

Travel between sites help team members to know each other, their culture and have informal communication and, to maintain trust among them. mitigation strategy alleviates the lack of face-to-face communication issue though socialization and create a feeling of team-ness. Social interaction among teams help to reduce interpersonal relationship issue and help to establish trust. Synchronous and asynchronous communication among team members help to reduce lack of face to face and informal communication issue. Trust-building takes time usually require frequent communication between parties. This strategy helps resolve any misunderstanding that might occur because of cultural and language diversity among dispersed team members. It also helps to develop team cohesion, interpersonal relationships among team members, which results in the improvement of informal communications among teams. Promote informal communication mitigation strategy is helpful to improve cohesion and interpersonal relationship among teams. In a GSD environment, informal communication can be done with the help of asynchronous

and synchronous communication channels. The remote team must communicate with each other and share their best practices, expertise, and knowledge by using efficient communication tools. This strategy is helpful to prevent the chance of schedule delays and to resolve cuticle problems which result in building trust among teams, to increase communication among team members, encourage effective use of groupware applications such as project management tools, wiki, Mendeley, drop box, Microsoft exchange etc.

In a nutshell, the current study contributes to the empirical evaluation of mitigation strategies for geographical distance risks. Moreover, the frameworks and their hypothesis has been tested that specify the impact of mitigation strategies on geographical distance risks. The survey is conducted from small and medium-size GSD organizations of Pakistan. These organizations made the project for the global market. Because of the distance involved, they face the same issues as other international GSD organizations face. The results of the study will be helpful to overcome the geographical distance risks that cause communication issues in the GSD environment, and it ultimately reduces the failure rate of a software project in Pakistan.

6. Conclusion and Future Work

GSD practice has been increasingly emerging in the software industry in recent years. Existing literature has observed that geographical distance is a crucial risk that hinders GSD projects' communication and leads projects toward failure. The geographic distance between dispersed teams cannot be reduced, but the potential effect of these risks can be minimized by applying different mitigation strategies. In a previous study, an SLR was conducted to identify geographical distance issues and their relevant mitigation strategies, and a conceptual framework was proposed but not empirically validated.

In this study, an empirical evaluation is performed. An online survey is conducted from the small and medium-sized GSD-based organizations of Pakistan to gather data and validate the hypothesis of the framework. As correlation among variables is pretty much high (>0.80), and collinearity is 7.041. So, to test the framework, we split it into eight frameworks and tested each mitigation strategy against its relevant issue. The finding of our study shows that all mitigation strategies have a significant impact on geographical distance issues with a p-value less than 0.01. So, we conclude that if organizations use mitigation strategies, the effect of geographical distance challenges will be reduced, and the study's finding is helpful in avoiding software project failure that occurs due to geographical distance risks.

In the future, an Analytic network process algorithm (ANP) can be used to prioritize the most critical strategy and geographical distance challenges. Moreover, a survey is conducted among small and medium-sized GSD organizations.

Like all other studies, this study also has a few limitations. The result of the study cannot

be generalized. Therefore, to generalize the results, it is recommended to conduct a similar study in some other countries. A few other electronic databases can be included in further studies to identify more issues and relevant mitigation strategies. The survey can also be conducted on the respondents of large GSD organizations in the future. In addition, the Analytical Hierarchical Process (AHP) approach may help the software industry prioritize the issues and mitigation strategies.

References

- [1] M. Shameem, C. Kumar, and B. Chandra, "Communication related issues in GSD: An exploratory study," 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA) pp. 1- 5 ,2015.
- [2] A. A. Khan, S. J. R. J. o. A. S. Basri, Engineering, and Technology, "A survey-based study on factors effecting communication in GSD," vol. 7, no. 7, pp. 1309-1317, 2014.
- [3] S. Mahmood, S. Anwer, M. Niazi, M. Alshayeb, I. J. I. Richardson, and S. Technology, "Key factors that influence task allocation in global software development," vol. 91, pp. 102 122, 2017.
- [4] J. Iqbal, R. Ahmad, M. H. N. B. M. Nasir, and M. A. J. J. o. I. T. Noor, "A framework to address communication issues during requirements engineering process for software development outsourcing," vol. 19, no. 3, pp. 845-859, 2018.
- [5] M. A. Shah, R. Hashim, A. A. Shah, and U. F. Khattak, "Communication management guidelines for software organizations in Pakistan with clients from Afghanistan," in IOP Conference Series: Materials Science and Engineering, vol. 160, no. 1, p. 012100, 2016.
- [6] M. D. KAHYA and C_. J. A.-e. B. T. O. D. S_ENELER, "A Literature Review on Challenges in Distributed Software Development," vol. 9, no. 35, pp. 159- 170, 2018.
- [7] J. Han, W. J. I. J. o. S. E. Jung, and I. Applications, "How geographic distribution affects development organizations: a survey on communication between developers," vol. 8, no. 6, pp. 241-252 ,2014.
- [8] B. H. Malik et al., "Geographical distance and communication challenges in global software development: A review," vol. 9, no. 5, 2018.
- [9] A. A. Khan, S. Basri, and P. Dominic, "Communication risks in GSD during RCM: Results from SLR," in 2014 International Conference on Computer and Information Sciences (ICCOINS), pp. 1-6, 2014.
- [10] N. G. de S_a Leit~ao J_unior, I. H. de Farias Junior, S. Marczak, R. Santos, F. Furtado, and H. P. de Moura, "Evaluation of a preliminary assessment method for identifying the maturity of communication in distributed software development," in Proceedings of the 2nd Workshop on Social, Human, and Economic Aspects of Software, pp. 12-18, 2017.
- [11] I. Qasim, M. Rashid, A. W. Khan, S. J. U. o. S. J. o. I. Khan, and C. Technology, "Prioritizing Geographical based Communication Oriented Risks and Associated Mitigation Strategies of Global Software Development," vol. 1, no. 1, pp. 25-34, 2017.

- [12] I. Nurdiani, R. Jabangwe, D. _Smite, and D. Damian,"Risk identification and risk mitigation instruments for global software development: Systematic review and survey results," in 2011 IEEE Sixth International Conference on Global Software Engineering Workshop, pp. 36-41,2011.
- [13] A. A. Khan, S. Basri, P. J. P.-S. Dominc, and B. Sciences, "A proposed framework for communication risks during RCM in GSD," vol. 129, pp. 496- 503, 2014.
- [14] Qureshi, Saim, Saif Ur Rehman Khan, and Javed Iqbal. "A Study on Mitigating the Communication and Coordination Challenges During Requirements Change Management in Global Software Development." IEEE Access 9 (2021),pp:88217-88242,2021.
- [15] Janjua, U. I., & Madni, T. M.,” Geographical Distance Issues and their Mitigation Strategies in GSD: A Systematic Literature Review towards Conceptual Framework. In 2021 4th International Conference on Computing & Information Sciences (ICIS), pp:1-6, 2021.
- [16] S. Morrison-Smith and J. J. S. A. S. Ruiz, “Challenges and barriers in virtual teams: a literature review,” vol. 2, pp. 1-33,2020.
- [17] U. I. Janjua, T. M. Madni, M. F. Cheema, and A. R. J. I. A. Shahid,” An empirical study to investigate the impact of communication issues in GSD in Pakistan’s IT industry,” vol. 7, pp. 171648-171672, 2019.
- [18] B. J. Babin and R. E. Anderson, Multivariate Data Analysis Joseph F . Hair Jr . William C . Black Seventh Edition, 2014.
- [19] W.-Y. Zhang, Z.-W. Wei, B.-H. Wang, and X.-P. Han, “Measuring mixing patterns in complex networks by Spearman rank correlation coefficient,” Phys. A, Stat. Mech. Appl., vol. 451, pp. 440–450, Jun. 2016.
- [20] Pearson's Correlation Coecient StatisticsSolutions:"[Online]:Available : [https : =www:statisticssolutions:com=pearsons correlation coecient=:](https://www.statisticssolutions.com/pearsons-correlation-coecient/)[Accessed : 24 june 2021]
- [21] N. Kock, User Manual: Version 6 . 0,” ScriptWarp Syst., pp.1122, 2018.
- [22] N. Kock and M. May-eld, PLS-based SEM Algorithms:The Good Neighbor Assumption, Collinearity, and Nonlinearity,” Inf. Manag. Bus. Rev., vol.7, no. 2, pp. 113 130, 2015.
- [23] Sweet, S. A., Grace-Martin, K. “Data Analysis With SPSS + Mysearchlab With Etext: A First Course in Applied Statistics”, 2011.

Appendix

Table 7: List of Mitigation Strategies and Risks

Risks	Mitigation Strategies
Lack of trust	<ul style="list-style-type: none"> -Travel between sites -Social interaction among team members maintain trust -Regular communication via synchronous devices -Encourage Frequent Communication -Promote informal Communication -Implement communication model -Encourage the use of effective communication tools and techniques
Lack of team cohesiveness	<ul style="list-style-type: none"> -Travel between sites -Encourage Frequent Communication
Lack of face-to-face meeting	<ul style="list-style-type: none"> -Travel between sites -Encourage synchronous communication -Encourage Frequent Communication -Promote informal Communication -Use groupware application -Use Effective communication tool
Lack of informal communication	<ul style="list-style-type: none"> -Promote informal interaction -Encourage frequent communication -Encourage effective communication through the adoption of tools and techniques -Use synchronous communication -USE asynchronous communication (instant messaging)
Lack of interpersonal relationship	<ul style="list-style-type: none"> -Social interaction among team members -Encourage frequent communication
Loss of communication Richness	<ul style="list-style-type: none"> -Travel between sites -Encourage frequent communication
Communication effort increase	<ul style="list-style-type: none"> -Encourage frequent communication -Use groupware application
Communication Frequency Reduced	<ul style="list-style-type: none"> -Encourage frequent communication -Use groupware applications

Comparative Analysis of Machine Learning techniques to Improve Software Defect Prediction

Muhammad Azam¹

Muhammad Nouman²

Ahsan Rehman Gill³

Abstract

One of the most active areas of research in the software engineering community is defect prediction. The gap between data mining and software engineering must be bridged to increase the rate of software success. Before the testing phase, software defect prediction predicts where these flaws will occur in the source code. Methods for predicting software defects are widely used to investigate the impact area in software using various techniques (clustering, statistical methods, neural networks, and machine learning models). The goal of this research is to examine various machine learning algorithms for predicting software defects. There have been many fault prediction techniques introduced, but no single technique or approach can be used for all types of datasets. To achieve maximum accuracy, different machine learning algorithms such as Bayesian Net, Logistic Regression, Multilayer Perceptron, Ruler Zero-R, J48, Lazy IBK, Support Vector Machine, Neural Networks, Random Forest, and Decision stump were used to uncover the largest subset of defects that could be predicted. This research concern is to find out defects using five NASA data sets JM1, CM1, KC1, KC2, and PC1. Logistic Regression has been shown to have the best output compared to others (93%).

Keyword: Software Defects, Machine Learning, Defect Prediction, SVM, Techniques.

1. Introduction

Defects designate unexpected performance of software system in return of user's given requirements. This abnormal behavior is usually found by the software tester in the phase of software testing marked as a defect. A software defect is also known as "Imperfection in the process of software development that usually causes software failure that could not meet the user's desired expectation." A defect is some deficiency or flaw in a software process or product. Because of an error, fault, or failure. The paradigm defines "error" as a human activity that promotes improper outcomes, and "defect" as a wrong choice that results in erroneous outcomes for a solution to the problem[1].

A software defect is about a condition where a software system or product could not meet software requirements or user's requirements, software having defects will be a failure.

¹University of Agriculture Faisalabad, Pakistan | writetoazamkhalid@gmail.com

²University of Agriculture Faisalabad, Pakistan | m.nouman909@gmail.com

³University of Agriculture Faisalabad, Pakistan | ahsanrehman41@gmail.com

Defects occur because of unusual or abnormal behavior of software or system. Unusual or abnormal behavior of software is directly proportional to software and as well it also affects customer requirements[2]. Since the major goal of testing the software in SDLC (software development life cycle) is to detect defects as soon as feasible, the team of software testing generally put the load on the testing phase to make sure that all defects are highlighted or found successfully, after the identification of defects and fix these defects in software testing phase by developers of software. Software defect existence influenced software reliability, software quality, and as well software maintenance price. It is impossible to achieve defect-free software, even the software testing process is put strictly just because most of the time defects are unseen. Stakeholders would ask the software testing team for forecasting software defects, so stakeholders could easily determine that the software is feasible and ready to deploy.

If the software is part defects this will be associated with some reason discussed below. Software Specification could also be wrong or not meet with user requirements either that could because of conflict requirements in the result software features be missing. It could be more complex to decide on the missing requirements that are not well explained or documented or might be poorly styled. It could not be considered that all requirements reflect wrongly[5]. Software developers could not be more competent for software projects just because of incomplete requirements. It could be the drawback of a project manager that the software development life cycle process will not follow by the project manager as needed in table 1.

Table 1: Defect Percentage

Software Development Phases	The defect may occur in a percentage
Requirement phase	20%
Designing phase	25%
Coding phase	35%
User Manual	12%
Bad Fixes	8%

- Software efforts are inclined to pay more observation on the following three fundamental issues in the software development life cycle:
- Software defect prediction from the huge amount of data[6].
- Software time estimation to ensure software reliability.
- Test software design, and test software process as well will affect the number of defects and density of software[7].

Defect prediction is an important activity to develop quality software. Because defect prediction precedes software deployment to reach user satisfaction and improves the

overall performance of the software. Identification of errors or defects earlier leads to sufficient allocation of resources that will because of reduction of time and cost as well to get a quality product. Therefore, the software defect prediction model participates in active responsibility for understanding the evaluation and improving the quality of software [8]. Different approaches have been planned to manage software fault prediction issues or problems. However, there are many techniques introduced in the literature review for fault prediction but there is no single approach for all datasets. Because it depends upon the nature of the dataset. Deciding which method should be used for fault prediction is a challenging activity. There is the most reliable method for defect prediction is Machine Learning[9]. To support a strategic distance from such disappointments in a software product, Defect prediction techniques (DPT) are performed in every stage of software development.

A. Software Quality Assurance:

According to past IT sectors and software firms, software quality is a prime object to focus on. Software defect prediction can straightforwardly impact software quality and achieved significant fame in a recent couple of years. Defective modules have a greater impact on software quality leading to cost, delivery time, and a lot higher maintenance costs [9]. Not being simply prepared to measure up to the assumptions on time and possibly speedy time is required, yet moreover, the ability to convey great quality programming things or infinitely better quality all the while is of most outrageous importance[10].

Hence, a lot of exploration is happening about how to further develop the item quality inside the compelled days open of the entire programming progression life cycle. Various ways exist for working on the overall idea of the item thusly made, for instance, better testing techniques, complete programmed testing works out, and early deformity expectation [11]. Thus, predicting software module whether a software entity contains defects can be helpful to improve the software quality. Therefore, quality is a key point that decides whether the software is according to a customer's need or process in which the software dataset under study is taken, and then pre-processing techniques are applied to data e.g., R, multilayer perceptron, k-neighbor nearest (KNN) and many more applied to retrieve information of defected data. In short, customer satisfaction is a key point for a successful project, for such a purpose we are going to research to find out defects earlier [12].

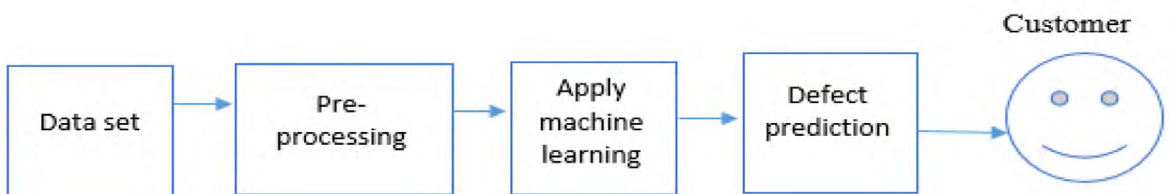


Figure 1: Generic Process of SDP

Software defect prediction achieved notable popularity in the last small number of years. Software Defect prediction directly affects software quality. Bad software modules have a solid impact on the software quality leading to cost overruns, delayed software completion timelines, and higher maintenance costs [13]. There are two basic approaches to Software Quality Assurance first one is defect detection and the second one is defect prevention. Defect prevention means avoiding upcoming defective activities as soon as earlier [12]. Defect prediction deals with existing defects. The approach to defect prevention is the process of improving software quality [14], and our research is concerned to improve software quality by predicting defects. Defect prevention activities are to find an error in software requirement planning, design the algorithm, and review the algorithm implementation [13]. The main object of defect prediction is to predict defects, bugs, or faults from software products and estimate the delivery quality and maintenance effort before the deployment process of the software product. The approach to defect prevention is the process of improving software quality [15].

The main research object idea is to explore the different machine learning algorithms to get maximum accuracy using feature selection for SDP. The prime aim of this research is to predict performance defects without overrunning the estimated cost and to find & analyzed which part of the software is more likely to have defects and deliver quality software.

The rest of this paper is as follows: Section 2 stands for related work. A machine learning algorithm is explained in section 3. The results of the experiments are discussed in section 4. In the end, section 5 concludes the paper and presents some future directions.

2. Literature Review

Most desired research zone of predicting defects using machine learning techniques, data metrics, and other techniques recently several research grounds started new projects. Various models and conclusions have been presented by scholars in different approaches. The investigation of more software defect prediction research papers published since the year 1990 to 2019 [16],[17].

Discussed machine learning algorithm for decision support which will achieve high precision and accuracy of decision support or decision-maker what was recommended. A deep understanding of the decisions making process was discussed. Two Methods of learning used implicit and explicit[18]. The researcher described that the non-symbolic knowledge provided better predictive accuracy in implicit learning and as well explicit produced symbolic knowledge which was a more comprehensible model[19]. In this paper researcher compared the comprehensibility and predictive accuracy of different machine

learning models (like explicit, implicit, and hybrid). These machine learning models are applied to several standard medical diagnostics, different electronic commerce, financial decision-making problems, and e-marketing. The best methods for every benchmark problem were different, but hybrid methods outperformed standard comprehensible methods, and as well ensemble methods often outperformed all other methods. Hoffding trees and their variants performed suitable for mobile computing, data streams, or big data and explained less accurate outcomes for these batch problems which have not a vast number of learning examples [20] discussed that quality, performance, and effectiveness of software attributes depended on the software defect prediction model. Eight NASA data sets were used to approach the right DP model. Characteristics of several software attributes are used to forecast whether software modules were defective? or non-defected? If proper attributes are not selected for the defect prediction model, the performance of the model will be decreased.

Therefore, [21] for The viability and execution of the imperfection forecast model, it is critical to choose suitable characteristics which could be utilized to construct a useful indicator model. The specialist proposed a quality determination cycle to distinguish damaged programming or to approach a legitimate SDP model. The trial result showed that the characterized approach gives a similarly productive arrangement of characteristics which expanded the presentation result, quality, and viability of the SDP model[21]. One ML classifier was used to improve the result of software defect prediction. In the future, the researchers look forward to integrating the performance of different ML classifiers to furtherly make an improved proposed approach. [26] proposed techniques that were intended to address particularly the blemished attributes of programming datasets, specifically, the absence of class-imbalanced information and marked examples, for imperfection expectation semi-directed approach of AI, were utilized. The scientist handled two basic issues of programming, comparably for programming deformity expectation, and proposed a semi-managed task-driven word reference AI strategy to foster both marked and unlabeled data completely. The exploratory outcome was performed on nine NASA datasets having subjective and quantitative information. This paper enhanced include extraction and the related classifier boundary, while different misclassification costs were investigated to further develop the classification exactness[22].

The experimental results demonstrate that our method outperforms several representative state-of-the-art defect prediction methods. [23] Used NASA's five data sets for software defect prediction. Different classifiers from the machine learning field: Naïve Bayes, neural networks, logistic regression, k-nearest neighbor, and support vector machine-implemented and evaluated on data sets, each classifier was evaluated on datasets from NASA's metrics. Personal implementations rather than WEKA were used for applying machine learning classifiers. Results of this research determined that all models could detect software defects using static features best model to results was:[24] Naïve Bayes

was overmatched in three datasets, SVM was overmatched by a k-nearest neighbor, and logistic regression for only one of the datasets. The information obtained from these machine learning classifiers was extremely effective for the continuous improvement of the software. In future work, research is defined as the choice of model is a difficult problem and requires more research, by increasing the number of the classifiers as well as a variety of datasets can uncover the underlying structure of software. [25] discussed how the data mining techniques used for defect prediction. Defective modules can be the cause of software failures, decrease customer satisfaction, increase development and maintenance costs and. The focus of their research was to find out the remaining defects from the software data set. Some data mining techniques.

Regression, clustering, classification, and association mining were used to predict flaws in the software and define how data mining techniques enrich the quality of software. [26] Conduct research to help software developers with issues of the undertaking, an attempt to detect defects from software has been made to make quality software. In this paper, the researcher formulates the software defect prediction problem as a classification task of machine learning[27]. It further assesses the impact of different outfit techniques to tackle the imperfection forecast issue. The course on programming imperfection forecast issues has been enunciated as an errand of classification and afterward, it reviews the impact of an assortment of group techniques on the AI strategy classification effectiveness. Specifically[22], researchers have derived a hybrid method of ensemble classification based on an over-sampled approach for software defect prediction in various imbalanced NASA datasets. The high unnecessary allocation of classes in datasets degrades the execution of classification approaches.

The proposed method has been derived based on the Synthetic Minority Over-Sampling Technique (SMOTE)[28][29]. In potential research work, the writer wants to include the verification process of the proposed method on various datasets. [30] proposed a model to forecast defects by the usage of software project metrics that were composed of a review of software design, product, review of source code per as per product deployed, and after defect validation was performed[30]. Linear regression (machine learning algorithm) was performed to selected metrics via software metrics only, software project metrics, and both. As a result, researchers declared that linear regression supplies the right correlation relationship between SD and predictors using both software and software metrics. One more thing proven in this research is to check out the suitability of the proposed analysis of regression to assemble an effective SDP model.

In a future direction, the researcher gives direction to the proposed model to fully automate, and standards rules could be weighted. So, the measurement could be depending not on the weight as well depend on the number of satisfied standard rules or violated rules. [32] Introduced the resample technique with three types of ensemble learners:

Boosting, Bagging, and Rotation Forest. These ensemble learners evaluated seven types of benchmark datasets of NASA. The researcher discussed Single Machine Linear Classifiers (Artificial Neural Network, Support Vector Machine, Locally Weighted Learning, Naïve Bayes, Decision Tree, Random Forest, J48 Decision Tree, Logistic Regression, and PART Algorithm) and Ensemble Machine Learning Classifiers (Bagging Techniques, Boosting algorithm and Rotation Forest). The researcher analyzed the accuracy and performance of three learners Boosting, Bagging, and Rotation Forest for the Software defect prediction dataset. The efficiency of performance, as a result, was loosed by using a Support vector machine algorithm with three homogenous ensemble methods and the researchers recommended that do not use Support Vector machine for defect prediction.

As mentioned above different researchers supply different solutions to overcome the error in the software. Some researcher uses software metric to overcome software error some use a different model to find the buggy module [37]. But no one can use early phase Software Defect Prediction. So, this is a unique research, and this research will provide great benefit to software engineering. During the literature review, it is found that there are some major techniques of machine learning like support vector machine, Bayesian networks, confusion metrics, clustering, Regression, Association rule, Bayesian Belief Network, Bayesian Belief network K-mean clustering, association rule, hybrid selection approach [38], genetic programming [38], k-mean clustering, genetic algorithm [38], static code matrix, automatic static analysis, association rule, and artificial neural networks are discussed to predict the faults.

3. Materials and Methods

Different approaches have been planned to manage software fault prediction issues or problems. However, there are many techniques introduced in the literature review for fault prediction but there is no single approach for all datasets. Because it depends upon the nature of the dataset. Deciding which method should be used for fault prediction is a challenging activity. There is the most reliable method for defect prediction is Machine Learning [31]. Feature selection involves evaluating better accuracy between input and desired variable using Minitab. WEKA machine learning tool is used for feature selection. For Statistical analysis mini tab was used to evaluate two-tail t-testing.

3.1 Feature Selection:

In the two study fields, ML and AI data analysis is the hottest topic for researchers. The feature selection technique determined an effective way to sort out various problems by extracting irrelevant or redundant data [32]. The major goal of feature selection is to improve the prediction performance based on accuracy precision and many more, provide

faster and cost-effective prediction within a time slot and provide a better understanding fundamental process of data generating [33].

The procedure of cutting the input variable that is not affected by the results is known as feature selection. In feature selection, researchers only select these features that key features of the dataset. This was desirable for researchers to reduce input variables to eliminate extra computational effort or cost of modeling, in some scenarios to improve the performance of the system or model. Feature selection involves evaluating the relationship between two or more variables. Selecting featured variables to have the strongest relationship between input variables and target variables Feature selection methods could be efficient, fastest, and effective.

As it is more challenging to use machine learning to select correct statistical measures for several types of datasets while it is performing filter-based feature selection. Data features used to train ML models have a huge influence on the accuracy or performance that could be better achieved. Irrelevant features or partially relevant features can negatively affect prediction model performance. Data set cleaning and feature selection should be the priority to design an effective model.

3.2 The algorithm is performed using machine learning:

Here are some algorithms discussed below which have been used for research results.

3.2.1 Logistic Regression

A logistic regression algorithm is used to predict the probability. It is used to prove the probability of a specific class for example pass/fail, win/lose, alive/dead, or sound/wiped out. This can be stretched out to display a few classes of occasions, for example, deciding if a picture has a feline, hound, lion, and so on. Each article found in the picture would be distributed a probability somewhere in the range of 0 and 1 and the aggregate added to one. It is a statistical technique and utilized for logistics purposes to display a binary (0 and 1 form) dependent variable albeit a lot of progressively complex augmentations exist [34].

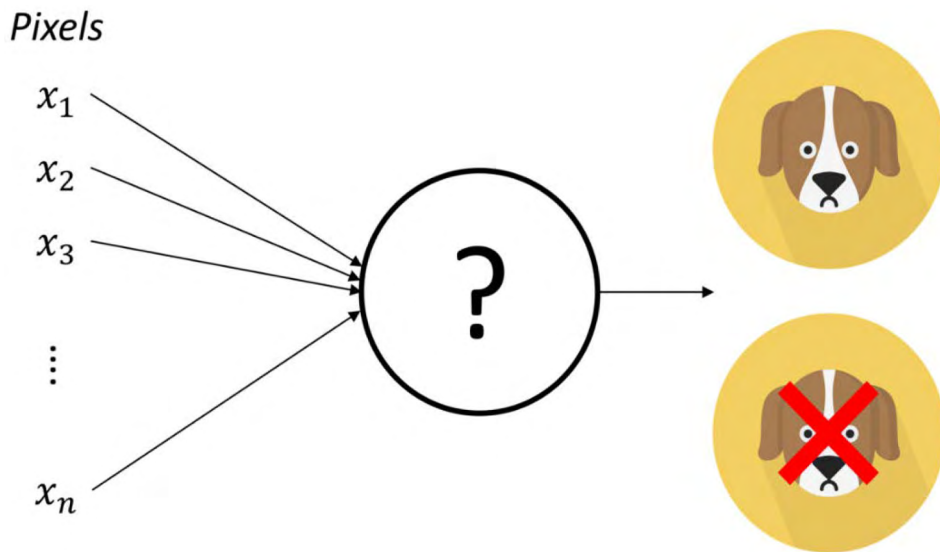


Figure 2: Logistic Regression

3.2.2 Bayes Net

Bayes Net uses to evaluate the probabilistic graphical model used by Bayesian inference for probability computations. Bayes Network determined model condition dependency by presenting conditions dependency of edges defined in the direct graph. Using Bayes Net, the researcher can perfect compact, variable factorization to join probability distribution to get advantages of conditional independence. Bayes network is used for prediction, anomaly detection, decision making in uncertainty situations, time series prediction, and many more [35].

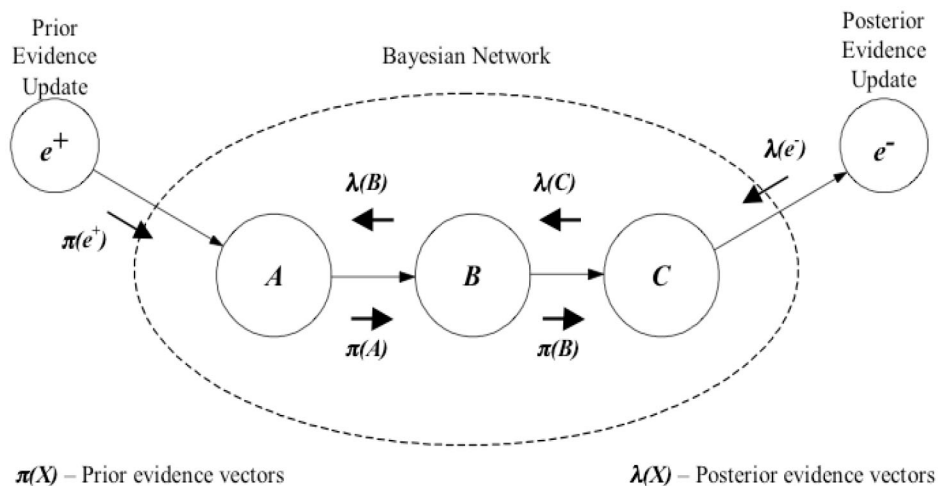


Figure 3. Bayes Net

3.2.3 Multilayer Perceptron (MLP)

It is a chain of perceptrons, systems of linear classifiers and it is a class of ANN. The term MLP is used abstrusely, here and there freely to allude to any feedforward ANN, now and again carefully to allude to a system made from different layers of perceptions. MLP are some of the times informally alluded to as “vanilla” neural systems, particularly when they have a single secreted layer. It consists of three layers of the node the 1st one is the input layer and the 2nd one is a hidden layer or unseen layer and the 3rd one is the output layer.

These three la

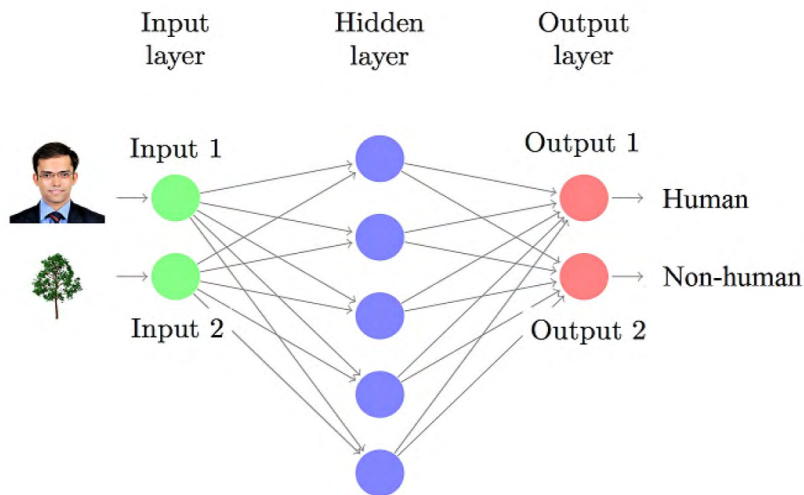


Figure 4: Multilayer Perceptron Layer Example

3.2.4. J48

J48 was used to build a decision tree derived by Ross Quinlan mentioned. J48 is the implementation of iterative Dichotomise 4 derived by the project team. J48 is a data mining tool for, the execution of C4.5 algorithms. It is an open-source Java implementation for deciding. Based on the input value, it generates tree-constructed data output. This theorem was developed by Ross Quinlan [36]. It holds continuous and discrete features. It is proper for bug prediction in all sizes of data sets. this rule is based on a learning classifier as a tree structure where every hub is a leaf.

3.3 Decision Stump

The decision stump ML algorithm has a one-level decision tree. DT is an exact technique for prediction. A decision stump build model to predict based on using a single input variable or feature. It is a well-ordered arrangement of if-Then instructions that can be more

minimal and, in this manner, more reasonable as compared to the decision tree. The choice to discover DT since it is less complex, and less computationally serious calculation than the decision tree method. It is the most straightforward method in ML. It outlines the data set with a DT which contains a similar number of qualities to the original data set [37].

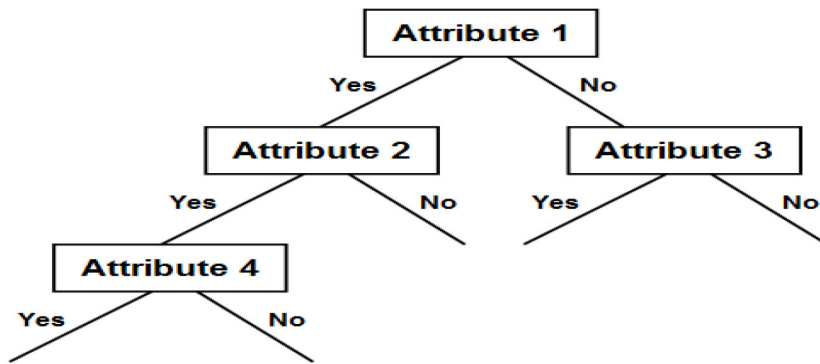


Figure 5: Decision Stump

3.3.1 Support Vector Machine:

SVM is supervised learning which could be used for regression, classification, and many more challenges. It arranges with individual perceptions or factors. In ML SVM are controlled learning models with related learning estimations that separate data utilized for relapse examination and grouping examination. Given a great deal of preparing models, each put aside as having a spot with both of two characterizations, an SVM preparing calculation fabricates a model that circulates new advisers for one class or the other, making it a non-probabilistic twofold straight classifier. An SVM model is a depiction of the models as spotlights on space, planned with the objective that the cases of the various classes are confined by a be normal. New models are then planned into that comparable space and expected to have a put with a class subject to the side of the opening on which they fall [38].

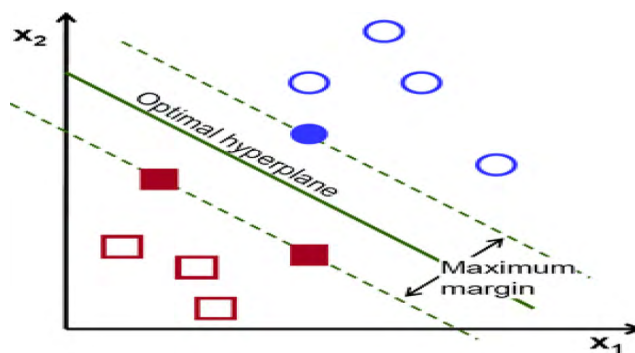


Figure 6: Support Vector Machine

3.3.2 Random forest

Data science presents a classification random forest algorithm. Random forest is the combination of DT, presented self-sufficiently with certain controlled change. It incorporates many trees, and the outcome is dependent on most of the precise yield (output) selected in the class. It is the greatest classifier for the huge data set. Each root node of the tree has a bootstrap sample data or information which is equivalent to the real data and each tree has different sample bootstrap. Utilizing the best split technique for factors or variables is arbitrarily chosen from input factors or variables.

Every tree is then developed to the most extreme degree conceivable without pruning. At the point when all trees are worked in the forest technique, new occurrences are connected to every one of the trees at that point voting process happens to choose the arrangement with the greatest votes as the original instance expectation [39].

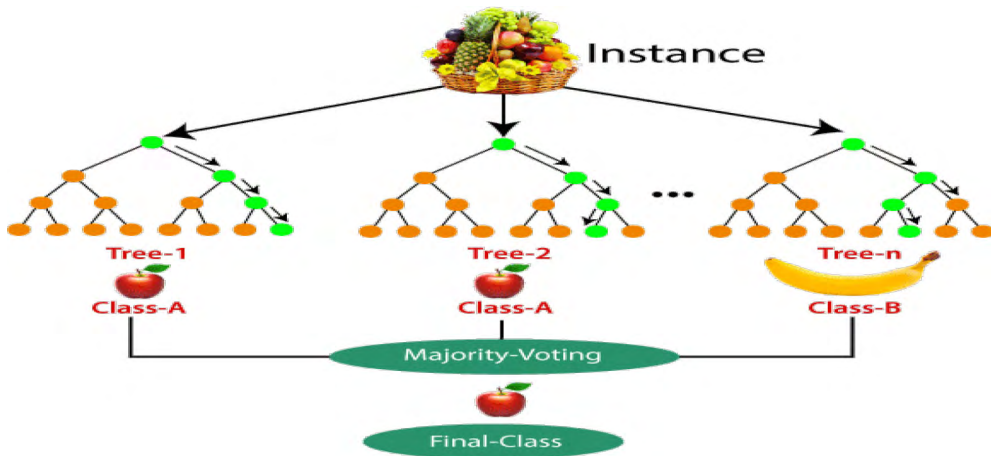


Figure 7: Random Forest

3.3.3 Lazy IBK

IBK is a fundamental algorithm family subset of classification. In k-nearest neighbor algorithm (KNN) is known as Lazy IBK (instance Based Learner). IBK is not useful to generate a model, instead, it is useful to build predictions for test instances within time. Distance is measured to find the k “closest” instance to make a prediction. It is an instance-based (IB) classifier. It varies from other IB learners in that it utilizes an entropy-based separation function [40]. It categorizes an instance by contrasting it with a database of pre-grouped models.

The key supposition that will be comparative instance will have comparative characterizations. The investigation lies in what way to characterize “comparative

instance” and “comparative classification.” The comparing segments of an IB are the separation work which decides how comparative two instances are, and the arrangement work which shows how case likenesses yield a last grouping for the advance or new instance. This strategy little bit slow to assess yet useful for expectation [41].

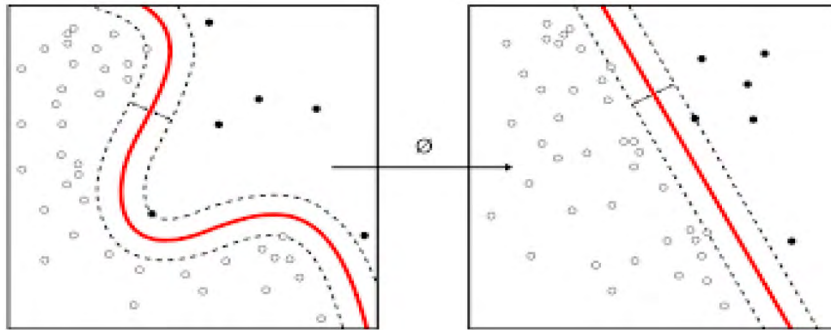


Figure 8: Lazy Inks

3.3.4 Ruler

Zero-R

Zero algorithms defeat One if the targeted distribution of data is limited and skewed available for predicting majority class, correct results basing a rule depend on a single attribute. Rule zero algorithm performed on nominal data type [42]. Rule Zero-R always exceeds baseline when it assesses the training data. In evaluating process training data may not be reflected by performance on independent test data.

3.4 Datasets

To promote the replication and verification for this research experiment. Publicly available benchmark datasets from the PROMISE Repository were used to get experimental results. Several datasets are available open-source and available on the internet. For this research, five datasets were obtained from NASA promise dataset repository CM1, JM1, KC1, KC2, and PC1 [43]. Table 1 supplies detail about each data set information like which language was used in the project, faulty instance, on-faulty instance, percentage of description Buggy, no of the attribute, and missing attribute et

Table 2 : Characteristic Of D Use

Project	Languages	# Of instance	Non-Faulty instance	% Of Des Buggy
CM1	C	498	499	9.83%
JM1	C	10885	8779	19.35%
KC1	C++	2109	1783	24.85%
KC2	C++	522	415 105	20.49%
PC1	NA	1109	1032	6.94%

Zero algorithms defeat One if the targeted distribution of data is limited and skewed available for predicting majority class, correct results basing a rule depend on a single attribute. Rule zero algorithm performed on nominal data type [10]. Rule Zero-R always exceeds baseline when it assesses the training data. In evaluating process training data may not be reflected by performance on independent test data.

For this research, six datasets were obtained from NASA promise dataset repository CM1, JM1, KC1, KC2, and PC1 [49]. Table 1 supplies detail about each data set information like which language was used in the project, faulty instance, on-faulty instance, percentage of description Buggy, no of the attribute, and missing attribute. Here is the data set shown below in a graphical form with several instances.

Here is the data set shown below in a graphical form with several instances.

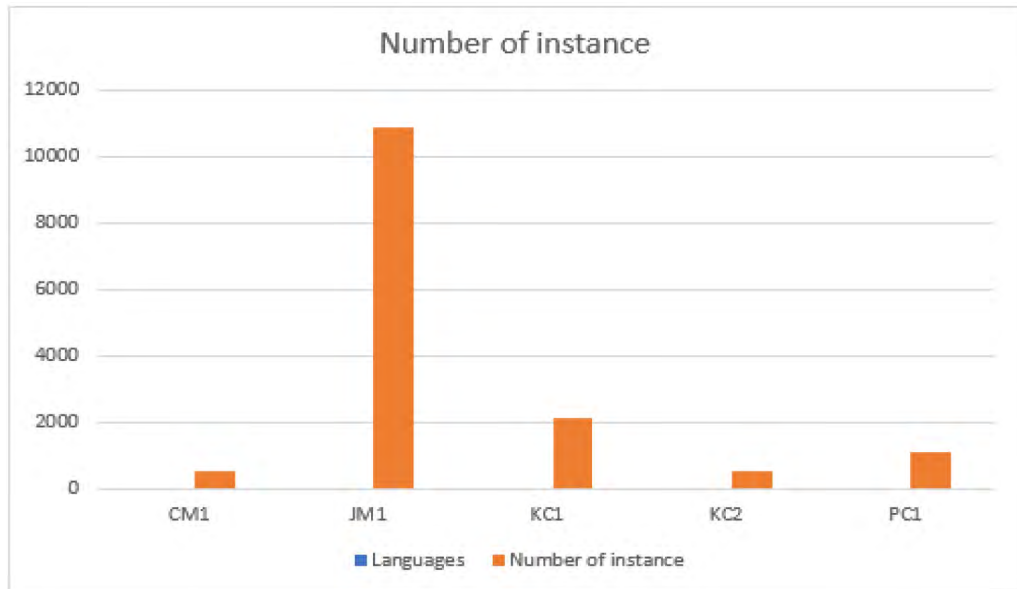


Figure 9: Number Of Instances

Here is the data set shown below in graphical form with the percentage of the buggy module.

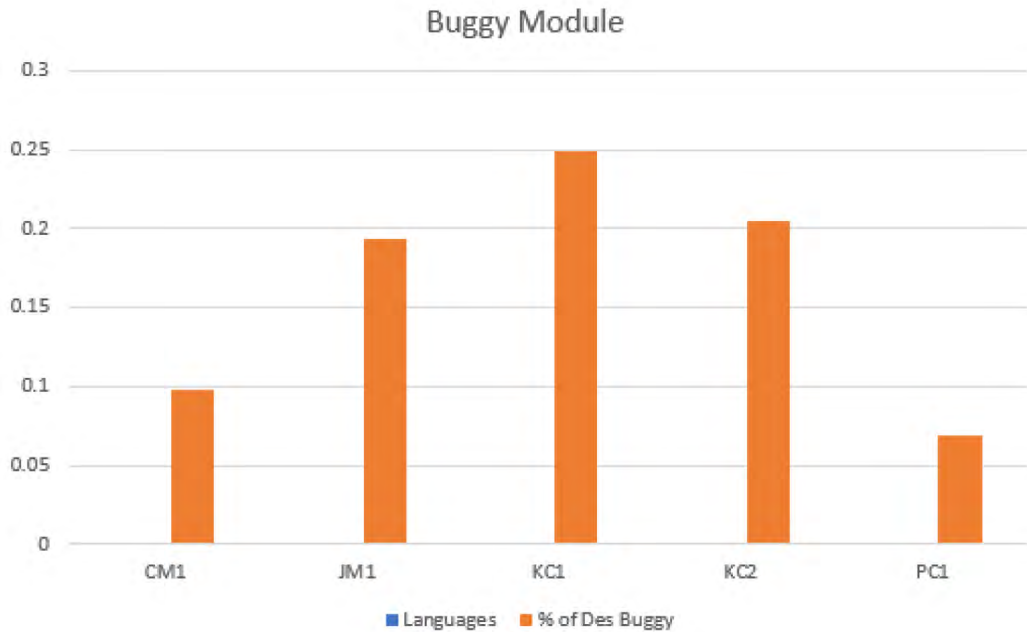


Figure 10: Buggies

4. Result And Discussion

4.1. NASA Repositor JM

JM1 accuracy with feature selection and without feature selection is in the following graph. It is clearly shown that feature selection accuracy is high as compared to without feature selection in figure 11.



Figure 11: JM1 Data Set Accuracy Graph

4.1.2. NASA Repositor CM

CM1 accuracy with feature selection and without feature selection is in the following graph. It is clearly shown that feature selection accuracy is high as compared to feature selection figure 12.



Figure 12: CM1 Data Set Accuracy Graph

4.1.3. NASA Repositor KC1

KC1 accuracy with feature selection and without feature selection is in the following graph 13. It is clearly shown that feature selection accuracy is high as compared to without feature selection.



Figure 13: KC1 Data Set Accuracy Graph

4.1.4. NASA Repositor KC2

KC2 accuracy with feature selection and without feature selection is in the following graph 14. It is clearly shown that feature selection accuracy is high as compared to without feature selection.



Figure 14: KC2 Data Set Accuracy Graph

4.1.5. NASA Repositor PC1

PC1 accuracy with feature selection and without feature selection is in the following graph 15. It is clearly shown that feature selection accuracy is high as compared to without feature selection.

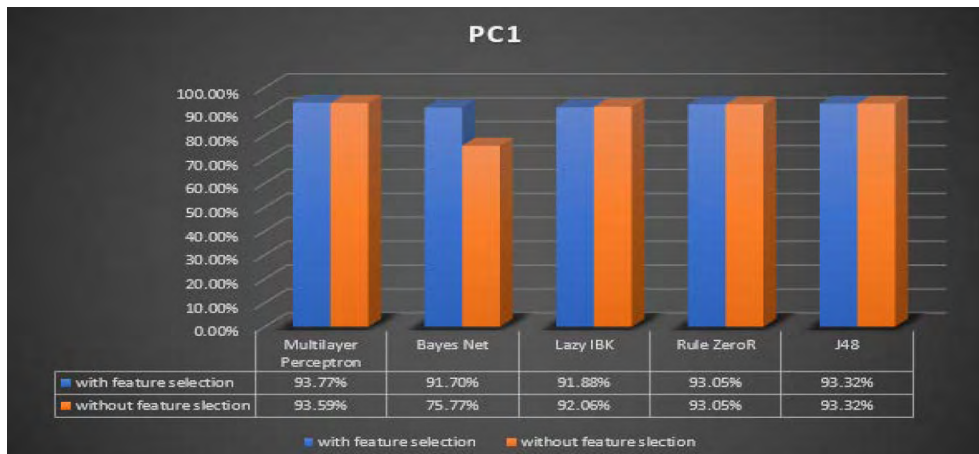


Figure 15: PC1 Data Set Accuracy Graph

To promote the replication and verification for this research experiment. Publicly available benchmark datasets from the PROMISE Repository were used to get an experimental result. Several datasets are available open-source and available on the internet. For this

research, Six datasets were obtained from NASA promise dataset repository CM1, JM1, KC1, KC2, and PC1 [43]. CM1 dataset used for prediction, JM1 dataset used for defect prediction KC1 used for prediction [44]–[46] KC2 dataset used in [43], [47], and PC1 dataset used in [48].

4.2. Results with Out Feature Selection

Accuracy performance without feature selection of 5 NASA datasets CM1, JM1, KC2, and PC1 is shown below tables 4.

Table 4.: Accuracy Table Feature Selection

Classifiers	CM1	JM1	KC2	PC1
Logistic Regression	73%	70%	78%	81%
Random Forest	83%	77%	82%	91%
Decision Stump	78%	71%	78%	87%
Support Vector Machine	75%	69%	79%	79%

All NASA dataset accuracy without feature selection is in the following graph. Here, in CM1, JM1, KC2, and PC1 datasets Random Forest is having the highest accuracy in figure 16. [49] used 10 cross-validation folds in which the dataset is divided into ten parts equally.

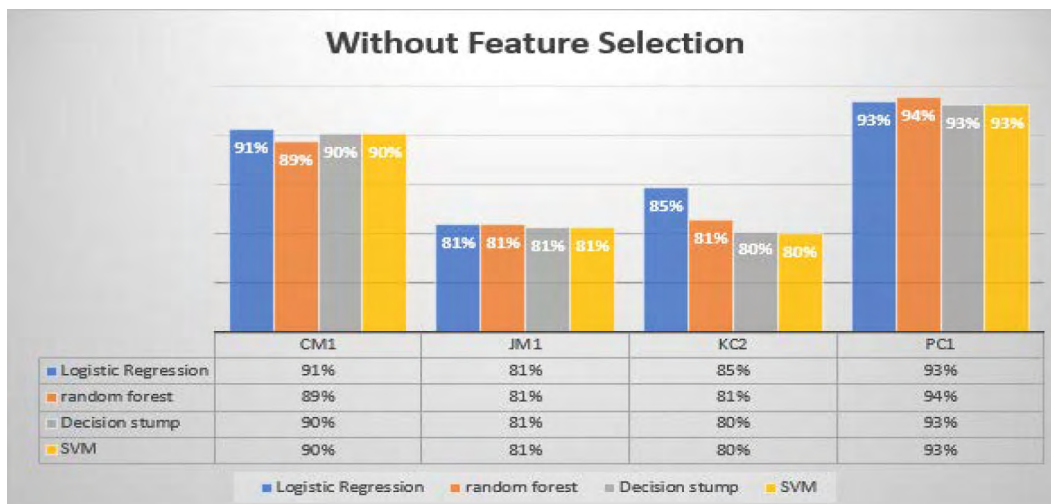


Figure 16: Accuracy Graph Feature Selection

4.1. Results with Feature Selection

Accuracy performance with feature selection of 5 NASA datasets CM1, JM1, KC2, and PC1 shown in below tables.

Table 5 : Accuracy Table With Feature

Classifiers	CM1	JM1	KC2	PC1
Logistic Regression	90.56%	80.94%	84.67%	93.32%
Random Forest	89.35%	80.92%	81.41%	93.86%
Decision Stump	90.16%	80.65%	80.26%	93.05%
Support Vector Machine	90.16%	80.66%	80.07%	93.05%

All NASA dataset accuracy feature selection is in the following graph. Here, in CM1, JM1 and KC2 datasets logistic regression are having the highest accuracy, as in the KC1 and PC1 data set Random Forest has the highest accuracy. In this research, thirty cross-validation folds in which the dataset is divided into 30 parts equally and test the dataset very closely and give a more accurate result. Overall PC1 accuracy is high by using all algorithms.

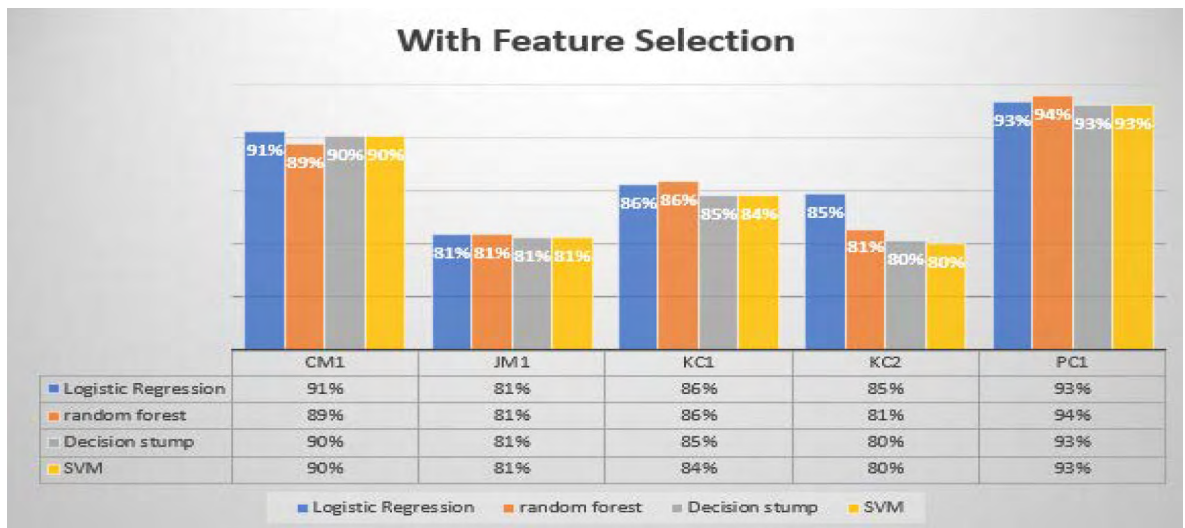


Figure 17: Accuracy Graph With Feature

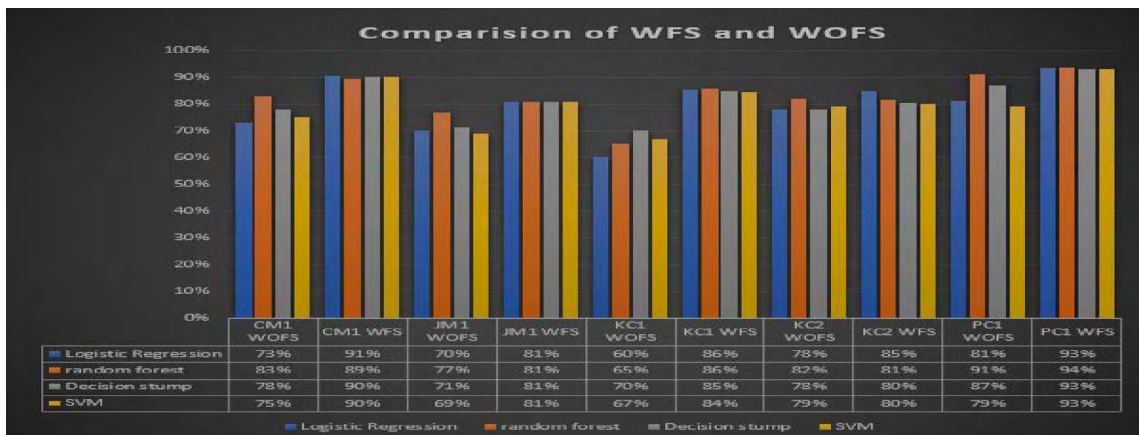
4.2. Accuracy comparison of WFS and WOFS

Accuracy performance with feature selection and without feature selection on five datasets CM1, JM1, KC1, KC2, and PC1 by applying logistic regression, random forest, and support vector machine shown below tables. Accuracy with feature selection is high as compared to accuracy without feature selection. In this research, thirty cross-validation folds in which the dataset is divided into 30 parts equally and test the dataset very closely and give a more accurate result. Here accuracy without feature selection is taken from.

Table 6: Accuracy Comparison WFS and WOFS

Data set	Classifiers	Accuracy WFS	Accuracy WOFS
JM1	Logistic Regression	80.94%	70%
	Random Forest	80.92%	77%
	Decision Stump	80.65%	71%
	Support vector Machine	80.66%	69%
CM1	Logistic Regression	90.56%	73%
	Random Forest	89.35%	83%
	Decision Stump	90.16%	78%
	Support Vector Machine	90.16%	75%
KC2	Logistic Regression	84.67%	78%
	Random Forest	81.41%	82%
	Decision Stump	80.26%	78%
	Support Vector Machine	80.07%	79%
PC1	Logistic Regression	93.32%	81%
	Random Forest	93.86%	91%
	Decision Stump	93.05%	87%
	Support Vector Machine	93.05%	79%

Here in the following table WFS= with feature selection WOFS=without feature selection. Here, in CM1, JM1, and KC2 datasets logistic regression is having highest accuracy, as in KC1 and PC1 data set random Forest has the highest accuracy. Overall PC1 accuracy is high by using all algorithms figure 18.

**Figure 18: Accuracy Comparison WFS and WOFS**

4.3. Results proved using statistics:

Two tail T-tests were applied using a mini tab to prove accuracy statically. The resulting

screenshot is shown below. For two-tail testing, two variables were used for testing accuracy with feature selection and accuracy without feature selection. This condition of p-value is shown below in the figure. Then H_0 will be accepted. But in this case, the p-value is 0.000 so H_0 is rejected. According to statistical decision accuracy with feature selection accuracy increased as compared to without feature selection. Using paired T-test statistical approach, it is proven that accuracy with feature selection is high.

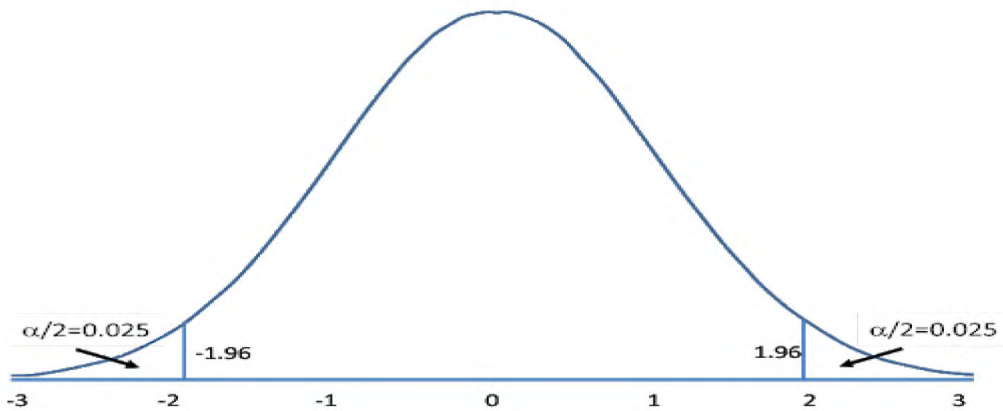


Figure 19: T-Test Result Graph

5. Conclusion

Software Defect perdition models aid to deal with these types of problems. Our research concerned was defected prediction by feature selection technique to get improvise accuracy results. This research result uncovers the largest subset of defects that could be predicted using above mentioned machine learning algorithm.

This paper's concern was to find out defects using Five NASA data sets JM1, CM1, KC1, KC2, and PC1. In this research machine learning algorithms Bayesian Net, Logistic regression, Multilayer perceptron, Ruler Zero, J48, Lazy IBK, Support Vector Machine, Neural Networks, Random Forest, Decision stump were used to perform feature selection to get maximum accuracy. Logistic Regression's highest accuracy founded at ninety-three% and the Bayesian Net averagely increase by an 8% accuracy rate using feature selection.

References

- [1] Al-Nusrat, Alaa, Fears Hamadeh, Mohammad Khorramshahr, Mahmoud Al-Ayoub, and Nahla Al-Dhahiri. [2019]. "Dynamic Detection of Software Defects Using Supervised Learning Techniques." *International Journal of Communication Networks and Information Security* 11(1):185–91.
- [2] Asaeda, Abdullah, and Mohammad Zubair Khan. [2019]. "Software Defect Prediction Using Supervised Machine Learning and Ensemble Techniques: A Comparative Study." *Journal of Software Engineering and Applications* 12(05):85–100. DOI: 10.4236/jsea.2019.125007.
- [3] Bernardo, Marco, Paolo Calcarine, and Lorenzo Donatello. [2019]. "Architecting Families of Software Systems with Process Algebras." *ACM Transactions on Software Engineering and Methodology* 11(4):386–426. DOI: 10.1145/606612.606614.
- [4] Boehm, Barry. [2019]. "A View of 20th and 21st Century Software Engineering." Pp. 12–29 in *Proceedings of the 28th international conference on Software engineering*. Shanghai China: ACM.
- [5] Bruckert, Remco R. [2018]. "Bruckert - Bayesian Nets in Weka." 23.
- [6] Bierman, Leo. 2017. "ST4_Method_Random_Forest." *Machine Learning* 45(1):5–32. DOI: 10.1017/CBO9781107415324.004.
- [7] Cai, Jia, Jiawei Luo, Shulgin Wang, and Sheng Yang. [2018]. "Feature Selection in Machine Learning: A New Perspective." *Neurocomputing* 300:70–79. DOI: 10.1016/j.neucom.2017.11.077.
- [8] Chen, Xiang, Yinzhou Mu, Key Liu, Zhan Qi Cui, and Chao Ni. 2021. "Revisiting Heterogeneous Defect Prediction Methods: How Far Are We?" *Information and Software Technology* 130:106441. DOI: 10.1016/j.infsof.2020.106441.
- [9] Dam, Hao Khan, Trang Pham, Shien Wee Ng, Tuyen Tran, John Grundy, Aditya Ghose, Takes Kim, and Chloe Kim. [2018]. "A Deep Tree-Based Model for Software Defect Prediction." *ArXiv:1802.00921 [Cs]*.
- [10] Devas Ena, Lakshmi, I. B. S. Hyderabad, and Lakshmi Devas Ena. [2018]. "Effectiveness Analysis of Xero, RIDOR and PART Classifiers for Credit Risk Appraisal Effectiveness Analysis of Xero, RIDOR and PART Classifiers for Credit Risk Appraisal." *International Journal of Advances in Computer Science and Technology (IJACST)* 3(11):6–11.
- [11] Esteves, Granderson, Eduardo Figueredo, Adriano Veloso, Markos Vigias, and Nivea Zaviana. [2020]. "Understanding Machine Learning Software Defect Predictions." *Automated Software Engineering* 27(3–4):369–92. DOI: 10.1007/s10515-020-00277-4.

- [12] Esteves, Granderson, Eduardo Figueredo, Adriano Veloso, Markos Vigias, and Nivea Zaviana. 2020. "Understanding Machine Learning Software Defect Predictions." *Automated Software Engineering* 27(3-4):369-92. DOI: 10.1007/s10515-020-00277-4.
- [13] Felix, Aquebogue Amara Chukwu, and Sai Peck Lee. [2017]. "Integrated Approach to Software Defect Prediction." *IEEE Access* 5:21524-47. DOI: 10.1109/ACCESS.2017.2759180.
- [14] Fu, Wei, and Tim Menzies. [2017]. "Revisiting Unsupervised Learning for Defect Prediction." *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering* 72-83. DOI: 10.1145/3106237.3106257.
- [15] Gruinard, Dominique, and Vlad Tarifa. [2017]. "Towards the Web of Things: Web Mashups for Embedded Devices." 8.
- [16] Gruinard, Dominique, Vlad Tarifa, and Erik Wilde. [2017]. "A Resource-Oriented Architecture for the Web of Things." Pp. 1-8 in *2010 Internet of Things (IoT)*. Tokyo, Japan: IEEE.
- [17] Humour, Wani, Mustafa Hammad, Mohammad Elnathan, and Fatima LaSharah. 2018. "Software Bug Prediction Using Machine Learning Approach." *International Journal of Advanced Computer Science and Applications* 9(2):78-83. DOI: 10.14569/ijacsa.2018.090212.
- [18] He, Peng, Bing Li, Xiao Liu, Jun Chen, and Yuta Ma. [2018]. "An Empirical Study on Software Defect Prediction with a Simplified Metric Set." *Information and Software Technology* 59:170-90. DOI: 10.1016/j.infsof.2014.11.006.
- [19] Herbold, Steffen, Alexander Tatsch, and Jens Grabowski. [2018]. "Global vs. Local Models for Cross-Project Defect Prediction: A Replication Study." *Empirical Software Engineering* 22(4):1866-1902. DOI: 10.1007/s10664-016-9468-y.
- [20] Herbold, Steffen, Alexander Tatsch, and Jens Grabowski. [2019]. "Correction of 'A Comparative Study to Benchmark Cross-Project Defect Prediction Approaches.'" *IEEE Transactions on Software Engineering* 45(6):632-36. DOI: 10.1109/TSE.2018.2790413.
- [21] Hutchison, David. 2014. "Future Data And." (2018). DOI: 10.1007/978-3-319-12778-1.
- [22] Jayanthi, R., and Lilly Florence. [2019]. "Software Defect Prediction Techniques Using Metrics Based on Neural Network Classifier." *Cluster Computing* 22(s1):77-88. DOI: 10.1007/s10586-018-1730-1.

- [23] Tiapride, Jaray's, Chakri Klan Tantithamthavorn, Hao Khan Dam, and John Grundy. [2022]. "An Empirical Study of Model-Agnostic Techniques for Defect Prediction Models." *IEEE Transactions on Software Engineering* 48(1):166–85. DOI: 10.1109/TSE.2020.2982385.
- [24] Kamila is, Andreas. [2019]. "A Lightweight Resource-Oriented Application Framework for Wireless Sensor Networks." Doi: 10.3929/ETHZ-A-005816888.
- [25] Kaur, Aras deep, Parminder S. Sandhu, and Manpreet Singh Brar. [2019]. "Early Software Fault Prediction Using Real-Time Defect Da" *2nd International Conference on Machine Vision, ICMV 2009* 242–45. DOI: 10.1109/ICMV.2009.54.
- [26] Kaur, Gaganjeet, and Amit Chhabra. [2014]. "Improved J48 Classification Algorithm for the Prediction of Diabetes." *International Journal of Computer Applications* 98(22):13–17. DOI: 10.5120/17314-7433.
- [27] Kondo, Masanari, Cor-Paul Beemer, Yasutaka Kamei, Ahmed E. Hassan, and Osamu Mizuno. [2019]. "The Impact of Feature Reduction Techniques on Defect Prediction Models." *Empirical Software Engineering* 24(4):1925–63. DOI: 10.1007/s10664-018-9679-5.
- [28] Lanza, Michele, Andrea Mocci, and Luca Panzanella. [2016]. "The Tragedy of Defect Prediction, Prince of Empirical Software Engineering Research." *IEEE Software* 33(6):102–5. DOI: 10.1109/MS.2016.156.
- [29] Taraji, Essam H., Mohammad Alshaya, and Lamoure Ghou. [2015]. "Software Defect Prediction Using Ensemble Learning on Selected Features." *Information and Software Technology* 58:388–402. DOI: 10.1016/j.infsof.2014.07.005.
- [30] Li, Jian, Panji He, Jiaming Zhu, and Michael R. Liu. [2017]. "Software Defect Prediction via Convolutional Neural Network." Pp. 318–28 in *2017 IEEE International Conference on Software Quality, Reliability and Security (QRS)*. Prague, Czech Republic: IEEE.
- [31] Li, Ning, Martin Shepperd, and Yuchen Guo. [2020]. "A Systematic Review of Unsupervised Learning Techniques for Software Defect Prediction." *ArXiv:1907.12027 [Cs]*.
- [32] Lyons, R. [2015]. "Approved for Public Release; Distribution Unlimited." 148.
- [33] Manjula, C., and Lilly Florence. [2019]. "Deep Neural Network Based Hybrid Approach for Software Defect Prediction Using Software Metrics." *Cluster Computing* 22:9847–63. DOI: 10.1007/s10586-018-1696-z.
- [34] MwanjeleMwagha, Solomon, Masinde Muthoni, and Peter Ochieng. [2014]. "Comparison of Nearest Neighbor (Ink), Regression by Discretization and Isotonic Regression Classification Algorithms for Precipitation Classes Prediction." *International Journal of Computer Applications* 96(21):44–48. Doi: 10.5120/16919-6729.

- [35] Naresh, E., Vijaya Kumar B. P, and Sahana P. Shankar. [2017]. "Comparative Analysis of the Various Data Mining Techniques for Defect Prediction Using the NASA MDP Datasets for Better Quality of the Software Product." 10(7):2005–17.
- [36] Neu, Lian. [2018]. "A Review of the Application of Logistic Regression in Educational Research: Common Issues, Implications, and Suggestions." *Educational Review* 00(00):1–27. DOI: 10.1080/00131911.2018.1483892.
- [37] Novakovic, Jasmine Đ., Alepine Valjavec, and Sinisa S. Ilic. [2016]. "EXPERIMENTAL STUDY OF USING THE K-NEAREST NEIGHBOUR CLASSIFIER EXPERIMENTAL STUDY OF USING THE K-NEAREST NEIGHBOUR CLASSIFIER WITH FILTER METHODS." (May 2018).
- [38] Pan, Cong, Minyan Lu, Biao Xu, and Hauling Gao. [2019]. "An Improved CNN Model for Within-Project Software Defect Prediction." *Applied Sciences* 9(10):2138. DOI: 10.3390/app9102138.
- [39] Parsons, Shaun, Rami Bassoon, Peter R. Lewis, and Xin Yao. [2019]. "Towards a Better Understanding of Self-Awareness and Self-Expression within Software Systems." 8.
- [40] Petrik, Jean, David Bowes, Tracy Hall, Bruce Christianson, and Nathan Badoo. [2016]. "The Jinx on the NASA Software Defect Data Sets." *ACM International Conference Proceeding Series* 01-03-June. DOI: 10.1145/2915970.2916007.
- [41] Singh, Kunwar P., Nikita Basant, and Shikha Gupta. [2016]. "Support Vector Machines in Water Quality Management." *Analytica Chemical Acta* 703(2):152–62. DOI: 10.1016/j.aca.2011.07.027.
- [42] Son, Le, Nakul Pritam, Manju Khari, Raghavendra Kumar, Pham Phuong, and Pham Thong. [2019]. "Empirical Study of Software Defect Prediction: A Systematic Mapping." *Symmetry* 11(2):212. DOI: 10.3390/sym11020212.
- [43] Son, Le, Nakul Pritam, Manju Khari, Raghavendra Kumar, Pham Phuong, and Pham Thong. [2019]. "Empirical Study of Software Defect Prediction: A Systematic Mapping." *Symmetry* 11(2):212. DOI: 10.3390/sym11020212.
- [44] Thota, Mahesh Kumar, Francis H. Sajin, and Gaultheria Rajesh. [2019]. "Survey on Software Defect Prediction Techniques." *International Journal of Applied Science and Engineering* 14.
- [45] Thota, Mahesh Kumar, Francis H. Sajin, and Gaultheria Rajesh. 2019. "Survey on Software Defect Prediction Techniques." *International Journal of Applied Science and Engineering* 14.
- [46] Wahoo, Rome Satria. [2015]. "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks." *Journal of Software Engineering* 1(1):16.

- [47] Wang, Po Wei, and Chi Jen Lin. [2014]. *Support Vector Machines*.
- [48] Wilde, Erik. [2007]. "Putting Things to REST." 14.
- [49] Xu, Zhou, Shuai Pang, Tao Zhang, Xia-Pu Luo, Jinn Liu, Yu-Tian Tang, Xiao Yu, and Lei Xu. [2019]. "Cross Project Defect Prediction via Balanced Distribution Adaptation Based Transfer Learning." *Journal of Computer Science and Technology* 34(5):1039–62. Doi: 10.1007/s11390-019-1959-z.

Statistical Analysis for the Traffic Police Activity: Nashville, Tennessee, USA

M. Y. Tufail^{1*}

S. Gul²

Abstract

Data Science is one of the fastest growing interdisciplinary field and has many applications in various disciplines. The actual motivation of data science came from John Tukey. In his seminal paper, in 1962, he presented the idea of data analysis which is now the field of data science. Several algorithms for data science related to statistical analysis have been developed and applied over variety of datasets since 1962. In this field, the significant development began with the aid of high performance computers that help to analyse a massive datasets. In this paper, we study the statistical analysis of the traffic stops in Nashville, Tennessee, USA for the year 2011–2021. Data is taken from the Stanford open policing project. Analysis is based on total number of 3071706 traffic stops. In this paper, we consider and investigate various aspects. This study comprises gender comparison (male vs female) and race comparison (black vs white) for different traffic offences. Complete findings and possible gaps are discussed in the conclusion.

Keyword: Data analysis, Statistical analysis, Traffic stops analysis, Traffic related social issues.

1. Introduction

Data Science is one of the fastest growing interdisciplinary field and has many applications[1–3]. John Tukey can be considered as the pioneer of Data analysis. In 1962, nearly 60 years ago, in his seminal paper [4] he published the idea of data analysis, that is now a field of data science [4, 5].

There have been many developments in data science since 1962. Data science is widely been used in various disciplines such as, social sciences [6–8], data engineering [9, 10], data mining [11, 12], predictive analytics [13, 14], machine learning [15, 16], image processing [17–20], data visualization [21, 22] and many more [23, 24].

^{1*}Department of Mathematics, NED University of Engineering & Technology, University Road, Karachi, 75270, Sindh, Pakistan | tufail@neduet.edu.pk

²Department of Mathematics, NED University of Engineering & Technology, University Road, Karachi, 75270, Sindh, Pakistan | sagul@neduet.edu.pk

The most significant boost in the field started due to the high-performance computers and the use of statistical analysis [25–27], that make this field inter-disciplinary and ease the gigantic calculations. Statistical analysis is almost used everywhere, whenever we deal with datasets [28–33]. Conclusions made in this paper heavily relies on statistical analysis.

In this paper, we apply statistical analysis to the data (during the year 2011 to year 2020) of the traffic stops by the police officers at Nashville, Tennessee, USA. Total number of traffic stops in the considered data is 3071706. Complete details of the data set can be found in the link <https://openpolicing.stanford.edu/data/>.

More than 20 million Americans are stopped each year for traffic violations [34]. Without any doubt, police stops is the most common form of interaction, the public has with the police around the world [35–38]. These activities and interactions help the police to achieve both traffic safety and crime control [39]. In recent years, the traffic stops data, is largely studied and analysed to understand the behavior of police and their interactions with the public, specially at traffic stops [40–43]. The law, prohibits, law enforcement agencies from stopping, detaining, or searching motorists when the stop is motivated solely based on the race, color, ethnicity, age, gender, or sexual orientation of drivers [6]. But incidents like George Floyd [44, 45] raise a very big question mark over police behavior, their act and their biased views towards specific race. These kind of incidents make the specific race vulnerable and cause the anger within the race towards other race. Further, we would also like to high-light the fact that (apparently) society is not gender biased. But we believe that the behavior of police towards the gender requires attention and statistical analysis.

This study comprises gender comparison (male vs female) and two race comparison (black vs white) for different traffic offences. Various aspects are considered and investigated. Complete findings and possible gaps are discussed in the conclusion.

List of desired variables (that are given in the form of columns in the dataset) are demonstrated in the Table 1 below:

Table 1: This table displays the variables and their corresponding details.

Column Names	Details of the variables Remarks
Date	Date of a traffic stop
Time	Time of a traffic stop
Subject race	The race of a subject
Subject sex	The gender of a driver
Violation	Eight different violations are presented in this column that include moving traffic violation, vehicle equipment violation, safety violation, registration, seatbelt violation, investigative stop, parking violation and child restraint.
Arrest made	It is a Boolean column [46, 47]. This columns has only two logical values, (i) 'True': this value indicates that arrest has been made and (ii) 'False': it Indicates that no arrest has been made due to traffic violation.
Outcome	This columns contains the information of the outcome corresponding to the respective traffic stop. Three distinct features for outcomes are available that include warning [48], citation [49] and arrest [50].
Contraband drugs	A Boolean columns consist upon two logical values. (i) 'True': it indicates that contraband drug is found and (ii) 'False: when contraband drug is not being found.
Contraband weapons	Analogous to the above explanation.
Frisk performed	A Boolean column analogously
Search conducted	A Boolean column analogous to other Boolean columns.

Table 2: Comparison of traffic violation between the male and female for Nashville, Tennessee, USA is given.

Male vs Female comparison for eight different categories of traffic violation				
Violations	Female	Female relevant ratio (individual values divided by total sum)	Male	Male relevant ratio
Moving traffic violation	633001	0.510	907881	0.50
Vehicle equipment violation	407379	0.330	585433	0.321
Safety violation	75168	0.060	110037	0.060
Registration	76994	0.062	107948	0.059
Seatbelt violation	33915	0.027	68932	0.038
Investigative stop	19275	0.015	36879	0.020
Parking violation	2992	0.002	4757	0.003
Child restraint	725	5.8×10^{-4}	390	2.14×10^{-4}
Total	1249449		1822257	

2. Statistical analysis

In this section we provide statistical analysis over five different examples using above mentioned dataset. The details are given below.

Example 1 (Traffic violation (Male vs Female)) In this warm-up example, we compare the count of eight different traffic violation committed by male and female for Nashville, Tennessee, USA during the year 2011–2020. We have found that nearly 50% of traffic violations are related to moving traffic violations (MTV) and this percentage is nearly the same for both male and female drivers. MTV are those traffic violations that occur when vehicle is in motion such as over speeding, stop sign violation, give way violation, driving under the influence of alcohol or drugs, hit and runs etc [51–55]. We have also found that the over all rate of traffic violation is higher in males than females. According to 2010 census, composition of male population is $\approx 48.5\%$ whereas; $\approx 51.5\%$ of female population [56, 57]. But whether or not the data is biased (with respect to number of drivers) would be an interesting future problem to address. Results are presented in Table 2 and graphical representation can be seen in Figure 1.

Table 2 suggests that most of the violations are related to moving traffic violations for both the genders. Nearly 50% violation involve MTV. Males are involved in 59% of the total violation whereas, females involvement is 41%. Both male and female are conscious when it comes to the child safety. Apparently, Table 2 indicates that males seems to be more concerned than female regarding child safety.

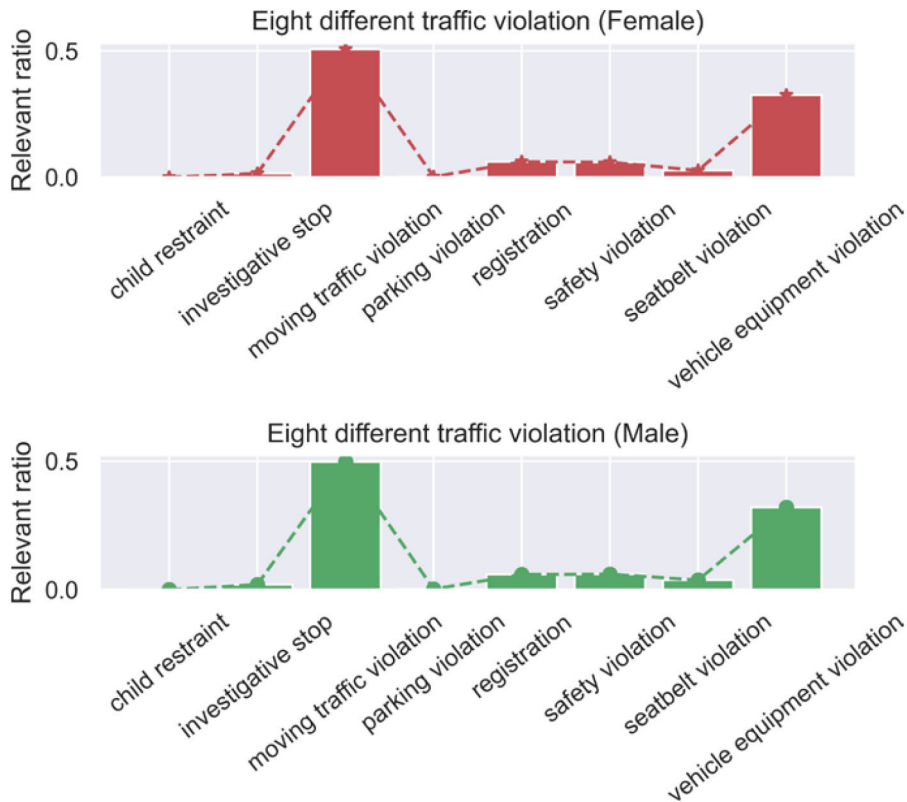


Figure 1: This graph represents the comparison of relevant ratios between male and female drivers among eight different categories of traffic violation whose details are given in Table 2. In all eight different categories, violation rate is nearly the same for both the genders. But for seatbelt violation, male ratio is slightly higher than female.

But, due to the limitation of current dataset, we have not investigated which gender transport their children mostly. It would be an interesting research problem and we aim to address it in future.

As the rate of moving traffic violation is higher than the rest of traffic offences and it covers nearly 50% of the total traffic violations. Therefore, we further explored this traffic offence in three different categories (with respect to the outcome/result of these traffic stops), i.e., warning, citation, arrest. Details are given in Table 3.

We further investigated the moving traffic violation among the different age groups. The frequency distribution for this classification is given in Table 4.

Table 4 indicates that the involvement of young driver in moving traffic violation is higher than the older people. This may be because most drivers are under 40.

Moving traffic violation: comparison between male and female		
Categories	Female	Male
Warning	0.6859	0.6712
Citation	0.3048	0.3103
Arrest	0.0093	0.0185

Table 3: Data represents the relevant ratios for the outcome due to moving traffic violations (MTV) for three different categories for female (total=632856) and male (total=907752) drivers out of 1540608 total MTV. The ratios are nearly similar for males and females. About 68% of stops for MTV result in a warning. This investigation also indicates that for MTV, the outcome is unbiased for gender.

Table 4: The frequency distribution for moving traffic violation corresponding to different age group.

Moving traffic violation VS Age group	
Age groups	Number of observations (x)
$10 \leq x \leq 20$	119391
$20 < x \leq 30$	485533
$30 < x \leq 40$	356318
$40 < x \leq 50$	269579
$50 < x \leq 60$	191360
$60 < x \leq 70$	87795
$70 < x \leq 80$	25097
$80 < x \leq 90$	5185
$90 < x \leq 100$	624
Total= $\sum x = 1540882$	

It is still an open problem because we are not too sure whether or not population has a similar number of people in all age groups (which is highly unlikely in unbiased data). Approximately 62.4% offences are caused by under 40 age group. The violations are keep decreasing with the maturity of a driver. Figure 2 illustrates this fact.

Example 2 (Search rate corresponding to each violation (Male vs Female)) In this example, we have calculated the mean search rate† among male and female drivers for

† Table 1 indicates that the 'search conducted' is a Boolean column (0: no search, 1: search is done). Mean search rate is simply an arithmetic mean ($\sum_{i=1}^N x_i/N$) for search corresponding to each violation.

eight different categories of traffic violations. We have found that the mean search rate for males are higher than female in all eight offences. Limitation of the data stops us to investigate the fact whether is data is being biased corresponding to specific gender. Investigative stops caused the highest mean search rate among all the violations for both the genders. Complete details are given in Table 5.

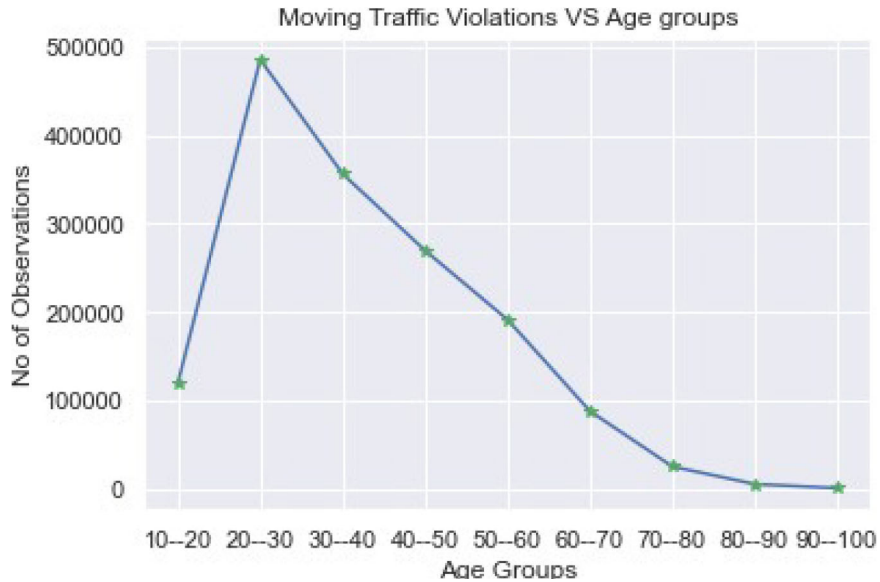


Figure 2: Comparison of moving traffic violations among different age groups. Young drivers are involved in more offences than the older drivers.

Table 5: This table shows the average values for search rate corresponding to each traffic violation for male and female. For all types of violations, it seems that the search rate is higher for males than for females.

Mean values for search rate corresponding to each traffic violation: Female vs Male		
Gender	Violations	Mean values
Female	Child restraint	0.030345
	Investigative stop	0.094319
	Moving traffic violation	0.019689
	Parking violation	0.024733
	Registration	0.025768
	Safety violation	0.023334
	Seatbelt violation	0.031903
	Vehicle equipment violation	0.024815
Male	Child restraint	0.071795
	Investigative stop	0.184089
	Moving traffic violation	0.046675
	Parking violation	0.047719
	Registration	0.055277
	Safety violation	0.048620
	Seatbelt violation	0.060277
	Vehicle equipment Violation	0.055152

Example 3 (Overall Search rate and Frisk rate (Male vs Female)) This example gives the details of overall search rate, frisk rate and arrest rate for male and female drivers. It is found that males are leading in all three categories in a given dataset (keeping the limitation of the data in account). Complete details are provided in Table 6.

Table 6: Male drivers are searched and arrested more than twice (approximately) as often as female drivers. Whereas, frisk rate among males are nearly three times higher than female drivers.

Overall mean Search rate, frisk rate and arrest rate between Male and Female		
Categories	Female	Male
Search rate	0.0235	0.0533
Frisk rate	0.008595	0.027683
Arrest rate	0.010633	0.019976

We have further investigated the overall all mean arrest rate for a period of 2011–2020 for a given 24 hours a day. '0' indicates the midnight, '12' represents the noon and '23'

states the 11:00 PM. Results are shown in Figure 3. It is found that the arrest rate is higher in overnight than any other hour of a given day.

Example 4 (Contraband Drugs and Contraband Weapons) In this example, we are trying to investigate whether or not the rate of contraband drugs and contraband weapons are increased over the past 10 years. Drug related stops are shown in Figure 4. Whereas, weapon related stops can be found in the Figure 5.

It is found that the drug related stops have kept increasing every year for the course of past ten years. However, weapon related stops have (continuously) declined during this period.

The comparison of drug related stops with the search rate for the period of past ten years is presented in Figure 6.

In this example, we have found that the drug related stops are increasing whereas, weapon related stops and search rate are decreasing. To check the validity of this claim, we further explore these trends. Non-parametric regressions [58, 59] results are given in the Figure 7 and the results of Modified Mann–Kendall [60] test are presented in Table 7.

Example 5 (Black race vs White race comparison corresponding to each violation) The motivation for this example came after the murder of George Floyd [44, 45, 63, 64]. This incident happened in Minneapolis, Minnesota, USA, dated: May 25, 2020. We investigate whether the police officers are biased for any race or they are neutral. Table 8 displays the comparison of black and white race drivers for eight different traffic violations.

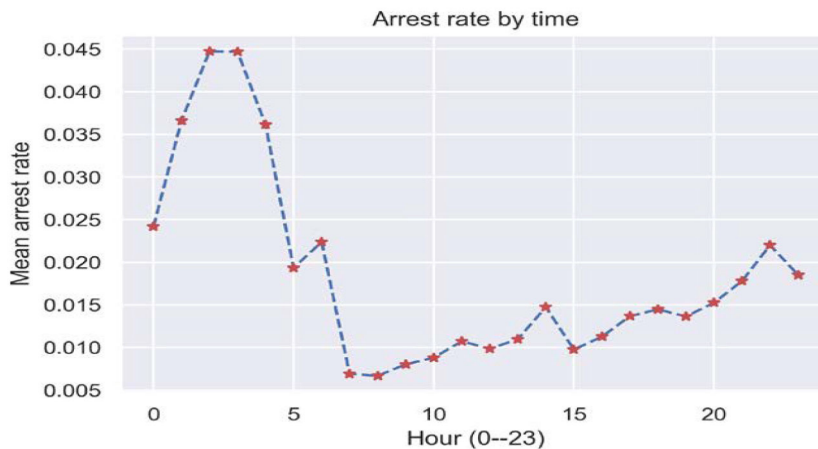


Figure 3: Graph indicates the hourly mean arrest rate for ten years period (2011–2020) for Nashville, Tennessee, USA. 0, 12, 23 indicate the midnight, noon and 11:00 PM respectively in a given day. The arrest rate has a significant spike overnight, and then dips in the early morning hours.

Table 7: This table confirms the actual trend with their respective parameters for Example 4. See [60–62] for more details.

Modified Mann-Kendall Test (at 5% level)		
Search rate	Drug rate	Weapon rate
0.0318	0.00067	0.1524
Decreasing	Increasing	No trend
-0.555	0.866	-0.377
-0.00068	0.018	-0.001
0.043	0.126	0.0196

Their respective relevant ratio can also be seen. Majority of the violations involve moving traffic violations (MTV) for both the races. Data suggests that there are total 2819799 traffic offences among black and white race. Out of which black people have committed 1158721 ($\approx 41\%$) offences. Whereas, white people are involved in 1661078 ($\approx 59\%$) offences. Apparently, this implies that white people committed more traffic offences than black race. But it is due to the

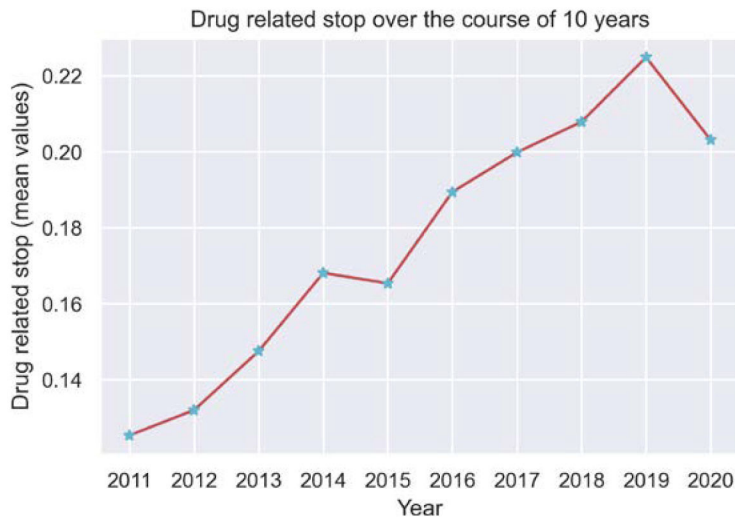


Figure 4: Graph indicates mean annual contraband drugs related traffic stops. Drug related stops are increasing every year. Surprisingly, The rate of drug related stops increased (nearly) doubled over the course of 10 years.



Figure 5: Graph indicates mean annual contraband weapons related traffic stop. Weapon related stops have decreased every year except during 2015–2017. The rate of weapon related stops decreased, and ratio is nearly one-seventh over the course of 10 years.

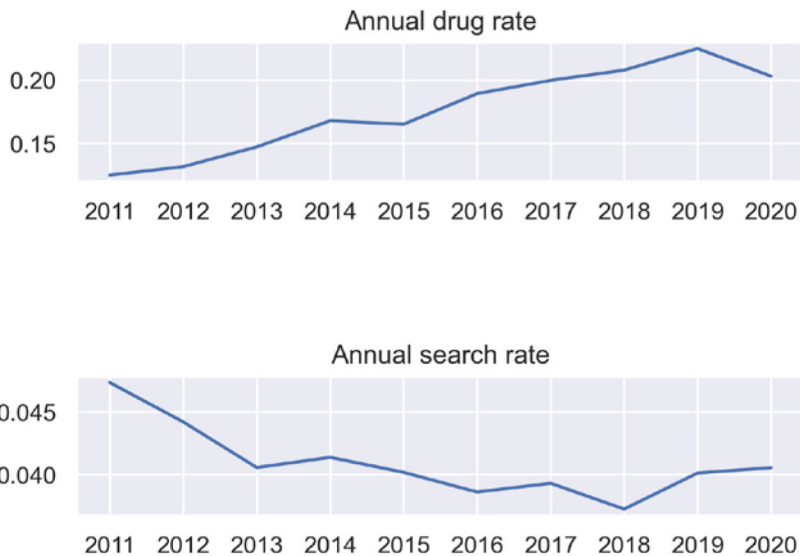


Figure 6: The rate of drug-related stops are continuously increasing. But surprisingly, the search rate is decreasing

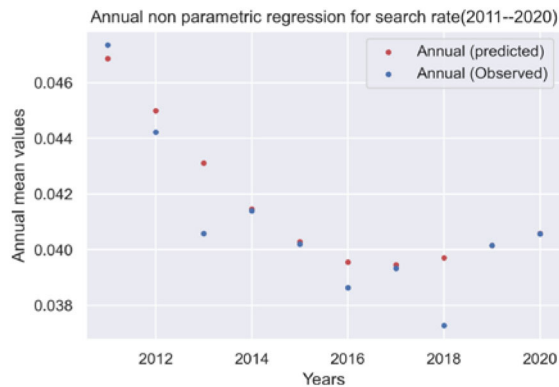
fact that the composition of white race is higher than the black race, i.e., $\approx 56.3\%$ versus $\approx 27.4\%$. (See [65, 66] for more details). The relevant ratios are plotted in Figure 8.

As we have found that the white race committed 59% traffic violations. Therefore we extend our investigation for three different categories that are search rate, frisk rate and arrest rate for both the races. Although (overall) white people committed more traffic offences than black but the mean search rate, frisk rate and arrest rate are higher in black race. Details can be found in Table 9.

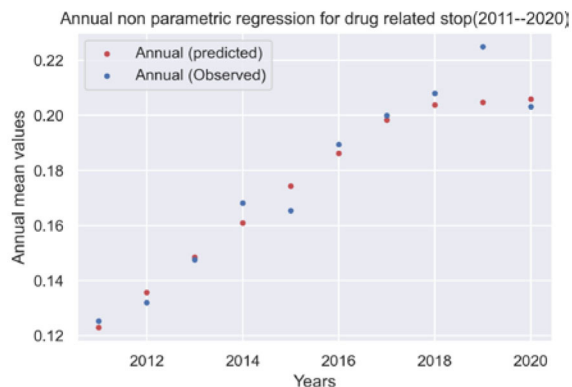
3. Conclusion

We have investigated the traffic police activity for Nashville, Tennessee, USA during 2011–2020. There are eight different traffic violations are investigated along with the outcomes of the offences (Example 1–3, 5). Example 4 contains the investigation of drug and weapon related stops. Overall findings are as below:

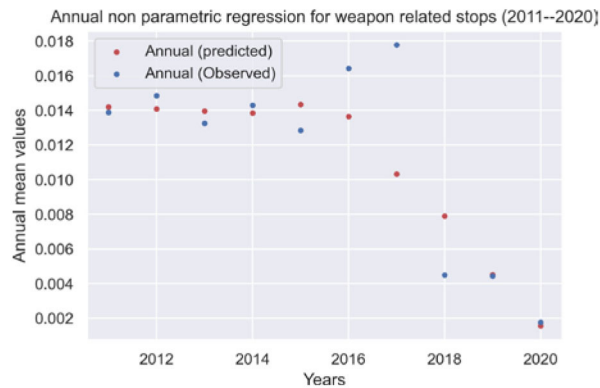
1. Over all males committed more traffic violations than female. But due to the limitation of given dataset, we have not investigated whether or not this dataset is biased.
2. The rate of moving traffic violation is higher than rest of remaining traffic violations.



(a) Search rate (2011–2020)



(b) Drugs rate (2011–2020)



(c) Weapon rate (2011–2020)

Figure 7: LOWESS regression for search rate, drug rate and weapon rate in Nashville, Tennessee during traffic stops are presented.

Table 8: This table shows the comparison between two different races (black vs white) corresponding to eight different categories of traffic violation. White race committed 1661078 traffic violation. Whereas, black race committed 1158721 traffic violation. Majority of the violations are due to MTV for both the races.

Comparison between black Race and white race corresponding to each traffic violation			
Race	Violation	Counts	Relevant ratio
Black violation	Child restraint	598	5.2×10^{-4}
	Investigative stop	25045	0.021614
	Moving traffic	532122	0.459232
	Parking violation	3719	0.003210
	Registration	74015	0.063876
	Safety violation	78708	0.067927
	Seatbelt violation	40954	0.035344
	Vehicle equipment violation	403560	0.348281
White	Child restraint	286	1.72×10^{-4}
	Investigative stop	25276	0.015217
	Moving traffic	880021	0.529789
	Parking violation	3480	0.002095
	Registration	98067	0.059038
	Safety violation	89684	0.053991
	Seatbelt violation	54939	0.033074
	Vehicle equipment violation	509325	0.306623

Table 9: This table illustrates the fact that (surprisingly) mean frisk rate, arrest rate and search rate are higher in black race drivers than white.

Overall mean Search rate, frisk rate and arrest rate between Black and White race		
Categories	Black	White
Search rate	0.058130	0.028535
Frisk rate	0.029388	0.012467
Arrest rate	0.022481	0.010810

3. Approximately 68% MTV results in a warning.

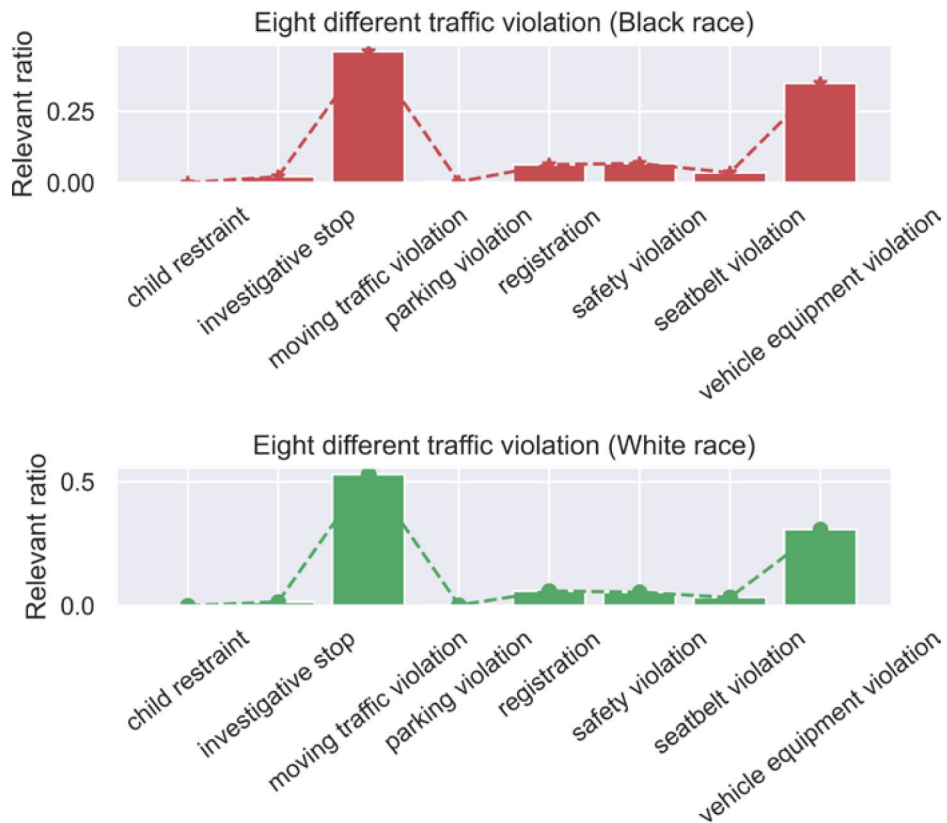


Figure 8: This graph represents the comparison between black and white race drivers among eight different categories of traffic violation. Majority of the violations are related to MTV for both the races. The relevant ratio for MTV is higher in white race than black. Whereas, vehicle equipment violation and safety violation are higher in black race. Overall white race committed more offences than black due to the fact that the composition of white race is higher than black.

4. Outcome of moving traffic violations do not indicate any kind of favoritism. Thus, the investigation concludes that for MTV, the outcome is unbiased for gender.
5. Overall search rate, frisk rate and arrest rate are higher in male's drivers than females.
6. We have found that the drug related stops are continuously increasing for the past ten years.
7. Weapon related stops indicated decreasing trend initially. But Mann-Kendal test confirms that there is no definite trend during these years.
8. We have also found that white race committed more traffic offences than black. But it is due to the fact that the composition of white race is higher than the black race. Nevertheless, the mean search rate, frisk rate and arrest rate are higher in black race.

Although we have found that males drivers committed more traffic violations than female drivers. But we have not investigated the gender of police officer. This could be our next task for the exploration. We would also like to explore the race of police officer and its impact over the decision. We have not investigated the effect of weather over a police officer. It would be very interesting question if we analyse the effect of weather over a decision.

References

- [1] Van Der Aalst, W.: Data Science in Action, pp. 3–23. Springer, London (2016)
- [2] Provost, F., Fawcett, T.: Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc., Sebastopol (2013)
- [3] Peyr'e, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. *Foundations and Trends@ in Machine Learning* 11(5-6), 355–607 (2019)
- [4] Tukey, J.W.: The future of data analysis. *The annals of mathematical statistics* 33(1), 1–67 (1962)
- [5] Tukey, J.W.: Available online. https://en.wikipedia.org/wiki/Data_science
- [6] McCabe, J.E., Kaminski, R.J., Boehme, H.M.: Racial profiling and ct motor vehicle stops: an observational study in three towns. *Police Practice and Research* 22(6), 1567–1584 (2021)
- [7] Scheb, J.M., Lyons, W., Wagers, K.A.: Race, gender, and age discrepancies in police motor vehicle stops in Knoxville, Tennessee: evidence of racially biased policing? *Police Practice and Research: An International Journal* 10(1), 75–87 (2009)
- [8] Renauer, B.C.: Neighborhood variation in police stops and searches: A test of consensus and conflict perspectives. *Police quarterly* 15(3), 219–240 (2012)
- [9] Gray, J., Shenoy, P.: Rules of thumb in data engineering. In: *Proceedings of 16th International Conference on Data Engineering (Cat. No. 00CB37073)*, pp. 3–10 (2000). IEEE
- [10] Birnholtz, J.P., Bietz, M.J.: Data at work: supporting sharing in science and engineering. In: *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, pp. 339–348 (2003)
- [11] Tan, P.-N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Pearson Education India, Noida (2016)
- [12] Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R.: *Data Mining for Scientific and Engineering Applications vol. 2*. Springer, Dordrecht (2013)
- [13] Eckerson, W.W.: Predictive analytics. *Extending the Value of Your Data Warehousing Investment*. TDWI Best Practices Report 1, 1–36 (2007)
- [14] Waller, M.A., Fawcett, S.E.: *Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management*. Wiley Online Library (2013)

- [15] Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. *Science* 349(6245), 255–260 (2015)
- [16] Goodfellow, I., Bengio, Y., Courville, A.: Machine learning basics. *Deep learning* 1(7), 98–164 (2016)
- [17] Murtagh, F., Starck, J.-L.: Image processing through multiscale analysis and measurement noise modeling. *Statistics and Computing* 10(2), 95–103 (2000)
- [18] Hill, D.L., Batchelor, P.G., Holden, M., Hawkes, D.J.: Medical image registration. *Physics in medicine & biology* 46(3), 1 (2001)
- [19] Tufail, M.Y.: Image registration under conformal diffeomorphisms: a thesis presented in partial fulfilment of the requirements for the degree of doctor of philosophy in mathematics at massey university, palmerston north, New Zealand. PhD thesis, Massey University (2017)
- [20] Marsland, S., McLachlan, R., Tufail, M.: Conformal image registration based on constrained optimization. *The ANZIAM Journal* 62(3), 235–255 (2020)
- [21] Chen, C.-h., Härdle, W.K., Unwin, A.: *Handbook of Data Visualization*. Springer, Berlin Heidelberg (2007)
- [22] Waskom, M.L.: Seaborn: statistical data visualization. *Journal of Open Source Software* 6(60), 3021 (2021)
- [23] Stedman, C.: Available online. <https://searchenterpriseai.techtarget.com/definition/data-science>
- [24] Hilbe, J.M.: A review of stata 9. 0. *The American Statistician* 59(4), 335–348 (2005)
- [25] Rzide-Gothenburg, L.: Statistics and the computer. *ICOTS*, 435–439 (1990)
- [26] Lunn, D.: Computer animation: A powerful way of teaching concept of probability and statistics. In: *Teaching of Statistics in the Computer Age. Proceedings of the 6th ISI Round Table Conference on Teaching of Statistics*, Chartwell Bratt Ltd., Bromley, pp. 114–125 (1985)
- [27] Diaconis, P., Efron, B.: Computer-intensive methods in statistics. *Scientific American* 248(5), 116–131 (1983)
- [28] Zammit-Mangion, A., Rougier, J.: Multi-scale process modelling and distributed computation for spatial data. *Statistics and Computing* 30(6), 1609–1627 (2020)
- [29] Kuttikkad, S., Chellappa, R.: Statistical modeling and analysis of high-resolution synthetic aperture radar images. *Statistics and Computing* 10(2), 133–145 (2000)

- [30] Schouten, B., Cigrang, M.: Remote access systems for statistical analysis of microdata. *Statistics and Computing* 13(4), 381–389 (2003)
- [31] Rafanelli, M.: Aggregate statistical data: models for their representation. *Statistics and Computing* 5(1), 3–24 (1995)
- [32] Takai, K.: Incomplete-data fisher scoring method with steplength adjustment. *Statistics and Computing* 30(4), 871–886 (2020)
- [33] Lyu, J., Gunasekaran, A.: Statistical analysis of a portable parallel non-linear programming algorithm. *Statistics and Computing* 4(4), 253–258 (1994)
- [34] Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., et al.: A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour* 4(7), 736–745 (2020)
- [35] Davis, E., Whyde, A., Langton, L.: Contacts between police and the public, 2015. US Department of Justice Office of Justice Programs Bureau of Justice Statistics Special Report, 1–33 (2018)
- [36] Langton, L., Durose, M.R., et al.: Police Behavior During Traffic and Street Stops, 2011. US Department of Justice, Office of Justice Programs, Bureau of Justice , USA (2013)

An Assessment for Understanding Student Behaviour by Applying Machine Learning Technique

Syeda Kainat Ahmed¹

Syed Mushhad M. Gilani*²

Sidra Sultan³

Abdur Rehman Riaz⁴

Muhammad Wasim Abbas⁵

Abstract

In the past, research have been carried out to study the behaviour of students as it is an important topic in psychology. Parents and teachers are concerned about their children's actions in class. Learning about students' behaviour in school is essential for teachers for their own development and growth of them. A class is composed of students with distinctive characteristics and capacities, where a few students are sharp and some are dull. In some cases, it becomes inconvenient for the educators to spot who is keeping the pace with them and who is falling behind. The proposed approach will allow the students to groom their personalities and overcome their shortcomings, and it will offer assistance to the instructor to identify which students require more consideration from them. To do that, we chose and applied the K-mean Clustering Algorithm. In other words, this research attempts to discover homogeneous subgroups inside the information. K means calculation is an iterative calculation that tries to tract the dataset into K pre-defined unmistakable non-overlapping subgroups (clusters), where each information point has a place as if it were one bunch. It tries to make the intra-cluster information focuses as comparable as conceivable, whereas also keeping the cluster as diverse (distant) as conceivable.

Keyword: Machine Learning, K-mean Clustering Algorithm, Behaviour Understanding, Student Psychology

1. Introduction

Child education plays an important role in the progression of any country [1]. In the future, they must serve and withhold the country. Thus, it is vital to understand the student's behavior to give them quality education in the best suitable environment, as education is

¹PMAS Arid Agriculture University, Rawalpindi | syedakainatahmed844@gmail.com

²The University of Agriculture, Faisalabad | mushhad@gmail.com (corresponding author)

³PMAS Arid Agriculture University, Rawalpindi | sidrasultan97@gmail.com

⁴PMAS Arid Agriculture University, Rawalpindi | abdurrehman.ar475@gmail.com

⁵Computer Science, University of Agriculture Wuhan, China | wasimabbas@whu.edu.cn

the right of every child [2]. It can only be possible under some expert guidance, which can teach them right from wrong, how to overcome their fears, and shortcomings, improve their grades, skills, effective learning methods, etc. [3]. A single class is composed of students with different characteristics and abilities, where some students are intelligent while some are less able. Sometimes it becomes difficult for the teacher to spot which student is picking up the pace of the teacher and which is falling behind [4].

It is observed that there is an increase in the number of students in educational institutes every year as compared to the previous year. To manage the increasing number of students, they are divided into different sections. The sections are formed generally and are not based on any particular characteristics of students, because, sometimes, it becomes difficult for the teachers to handle the large strength of a different variety of students. Some students are shy while some are confident, some are disorderly while some are obedient, and some are dull while some are intelligent. There is a difference in the interaction of students with other students and teachers as well. So, in this project, a system is designed that predicts the behavior of students as given by [5]. For this, firstly a survey was designed containing several questions that were filled in both by the students and teachers. In this, students with similar characteristics are grouped in one section and a teacher is assigned to them according to their behavior. This helped the teachers to make sure that everybody was moving along at the same pace while delivering lectures, the school (staff, administration) to improve their results, parents to keep an eye on their children's progress, and also the students to groom their personalities and overcome their shortcomings [6].

This research work allows the students to be grouped according to their capabilities, which helps the teachers to learn which students require more attention from them. It helps both teachers and students to deliver their lectures more effectively and to improve their grades and skills, respectively. For this purpose, data was gathered from different schools, then students were divided into different categories according to their behavior through the application of Machine Learning (ML) techniques. Students with similar behavior and attributes were assigned to the same section. Lastly, the teacher whose personality matched with those students was made responsible for their teaching. This removed the communication barriers between the teacher and students and provided a friendly class atmosphere for students to express themselves. Parents also had access to their children's progress reports. Moreover, the school administration had a record of teachers' and students' reports [7]. In section I, Introduction has been discussed. Section II deals with the related research while Section III describes the proposed system and section IV results are shown.

2. Related Work

The research was performed to study the effects of hospitalization on children [8]. The online

survey was taken from the parents and its outcome revealed the types of mental pressure and diseases faced by the children. The studies also suggested some remedies for these types of disorders. Another experiment was conducted in a school to observe the behavior of children while evacuating in non-emergency situations without any guidance using the Cellular Automata CA model [9]. It was perceived that children in groups were taking longer time than those who were walking alone. In the end, they matched the results with their simulated results to ensure the correctness and precision of their designed system [10].

Zheng, Jiang and Shen [11] proposed a system that can spot attitudes with low-resolution and extreme impediments to help teachers, they can improve their teaching quality. They built a large-scale data set. An improved Region-Based Convolutional Neural Network (R-CNN) network was presented to discover student behaviour in the real classroom. Online Hard Example Mining (OHEM) was used for class imbalance issues. Fan et al. [12] describe a system that can remodel the activity occasions of the examinee about signaling data obtained by Kinect. It examines the attributes of activity occasions from period and recurrence proportions. The misconduct or disobedience of the examinee was also found in the studies. They did it by information procurement, pre-preprocessing module, occasion distinguishing module, misconduct observing tool, and track record constitute [13].

The research in [14] detects the behavior of the students while doing the class tasks and activities by using a round-robin coding strategy, regression tree, and observing students with the naked eye [15]. It suggested that if students are looking at the teacher, they are on-task. Otherwise, they are off-task. Some of the reasons for off-task behavior are self-intrusion, peer diversion, natural diversion supplies, strolling, or other self-distractions. Singh [16] diagnosed Attention Deficit Hyperactivity Disorder (ADHD) in children that experienced loss of self-control because of cognitive disabilities and unfriendly habitat and observed their behavior under different circumstances. It was discovered that the children can control themselves if they are not bothered unnecessarily. It was done by grouping children into 3 and then interviewing them for one hour with four different female researchers. Some questions were also asked by the parents because parental behavior also influences children's behavior [17].

In another method, 51 children ages ranging from 4 to 6 years were interviewed (qualitative approach) and asked what happiness is to them and what and who makes them happy [18]. The studies show that, not feeling tired or lazy and having positive affiliations, playing, learning, writing, drawing, helping their mothers, reading comics, watching TV and sports make them happy. Relationships with people in the family, friends, and their toys are also a source of happiness for them [19]. Lima et al. [20] discussed causes and remedies for childhood depression. They did a qualitative analysis of 180 articles, out of which 25 were in their domain. Out of those 25 articles, data was extracted and arranged in compiled form. The outcome of the research was based on etiology, diagnosis, prevention, prognosis, and treatment [21].

Safaei and Youzbashi [22] held a correlated comparative study, a multistage method, and used SPSS Software version 25.0 for data analysis. Their research included 140 people, out of which 80 were girls and 60 were boys, and their ages ranged from 8 to 10 years for the contrast of seriousness of the fixation and working memory in children with the obsessive-compulsive disorder and healthy children. The outcomes indicated that there was a huge connection between working memory and habitual issues in girls and boys [23]. The figures for mean and Standard Deviation (SD) in youngsters and patients were (177.24+-11.02) and (171.11+-8.08) respectively.

Kessels and Heyder [24] inspected the mental convenience of locks engaging in disruptive behavior for low-achieving students from an attributional point of view. In their experiment, 178 ninth-grade students were selected, and targeted the those students, who displayed disruptive behavior. They connected multilevel examination while testing for mediation effects. It was also discovered that those students were more popular but not liked personally. The research also included the comprehension of troublesome behavior in class as an endeavor of children showing unsatisfactory results, to inspire face-saving attributions and improve their peer status [25]. A taxonomy of the literature review is shown in Figure 1. Table I shows the previous literature and also discuss their finding and limitation.

Table 1: Comparison table of related Studies

State of the art Approaches	Findings	Limitations
Online questionnaire for parents about behaviour of children (Jiao et al., 2020).	Parents were instructed to, spend time with their children, and develop an interest in music for mental relief and comfort.	Only suggestions were given no algorithm was applied
Cellular Automata (Chen et al., 2019).	Grouped children took more time for evacuation A path with fewer obstacles regardless of distance was chosen	Behaviour of children under guidance was also not detected
OHEM combined with R-CNN is used to detect student behaviour (Zheng, Jiang and Shen, 2020).	The proposed system detects more behaviors with low-resolution and severe occlusion and helps teachers to improve teaching quality.	Behaviour of Students with selected poses was detected
An approach consists of four components data acquisition and pre-processing module event identification module misbehaviour monitoring engine record module (Fan et al., 2016).	Their approach rebuilds the action events of the examinee in terms of gesture and then these gestures are used to detect misbehaviour	There exists a possibility that the input data may comprise garbage value.
51 kids ranging from 4 to 6 years were interviewed (Izzaty, 2018).	55% of children said not feeling lazy or tired is happiness	No objective analysis was carried out. The sample size was too small (size=51)

Multistage cluster sampling method (Safaei and Youzbashi, 2020).

Healthy groups had higher working memory and lesser severity of obsession compared to patients

A very small number of students were taken for tests and all were of adult age.

Multilevel analysis while testing for mediation effects (Kessels and Heyder, 2020).

Disruptive behaviour causes a lack of effort instead of a lack of interest.

There is no information that whether the disruption was conscious or not.

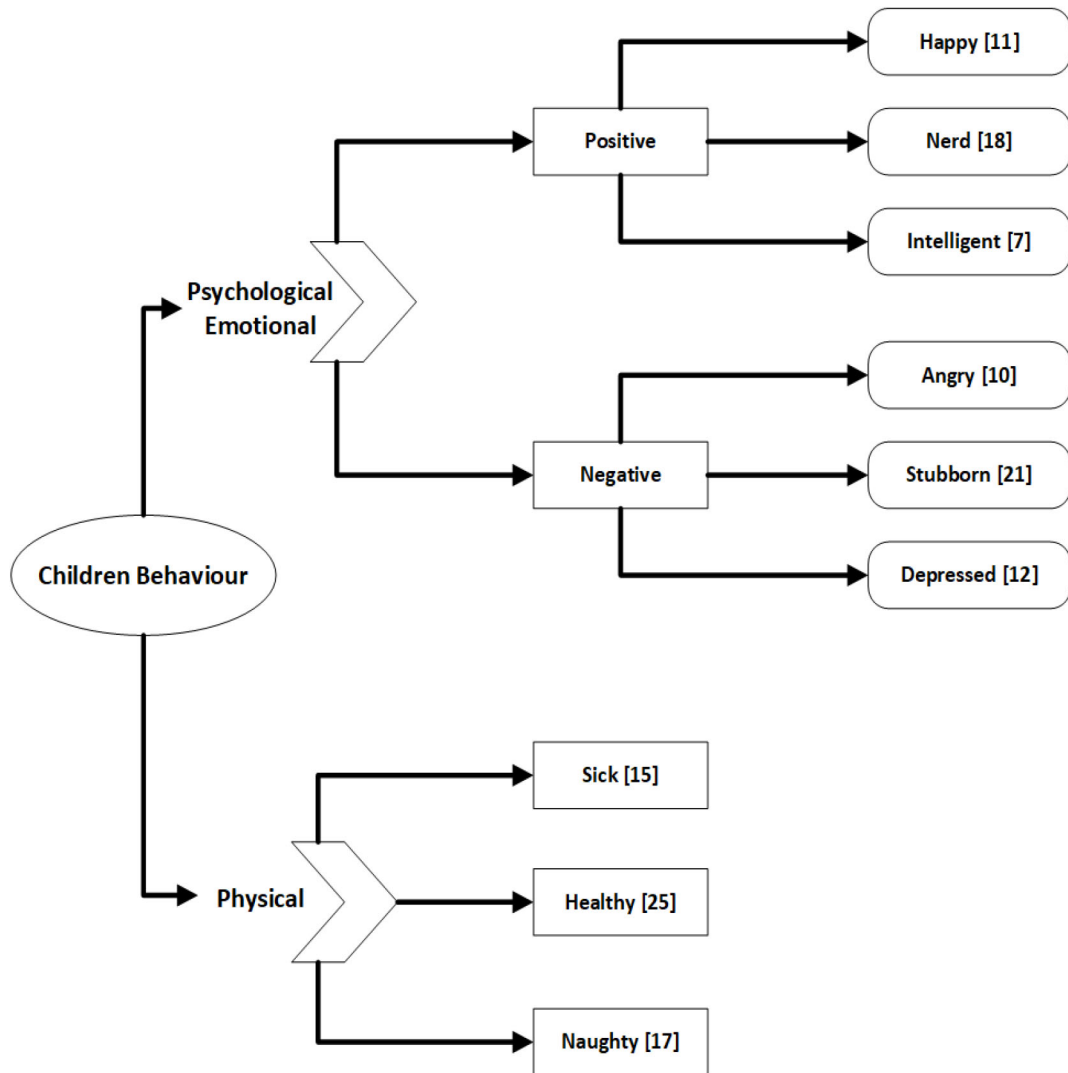


Figure 1: Taxonomy of child behavior interconnected examinations

3. System Model

The proposed model is supported to clarify the kinds of data that can emerge from a thematic approach for analysis that provides simple guidance on the opportunities and limitations of such data. The proposed model consists of five stages, as illustrated in Figure 2.

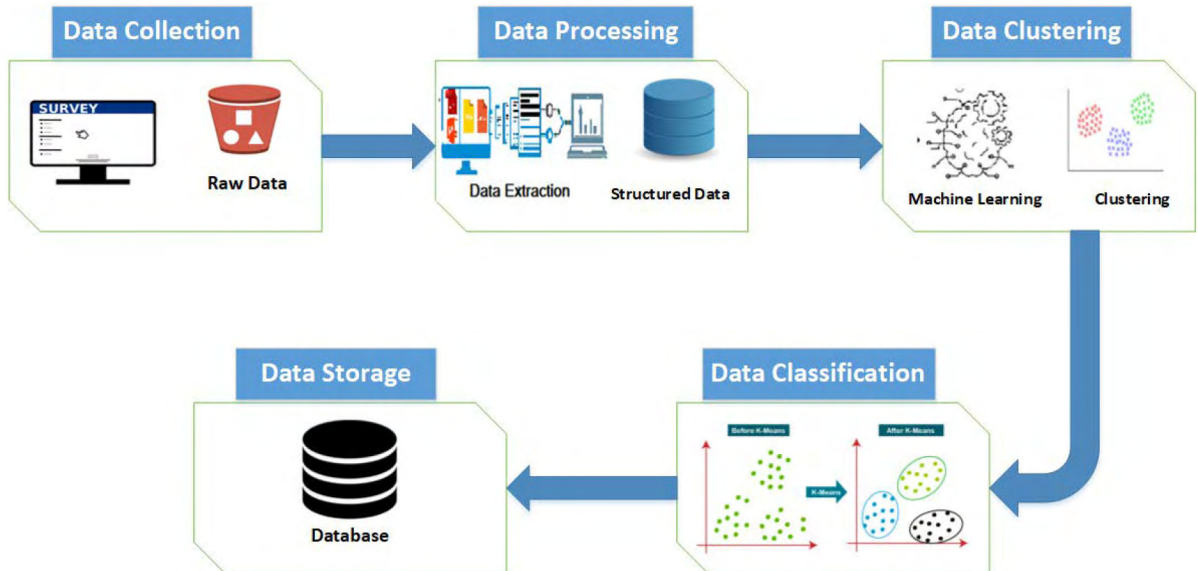


Figure 2: Proposed system

A. Data Collection

The first stage of the system model is data collection, in which a survey is conducted to store initial raw data. We designed diverse kinds of questions for students and asked them to rate themselves from 1 to 5. Where 1 was for the lowest and 5 was for the highest rank. The collected data is arranged in the form of charts and spreadsheet files, as details are presented in Figure 3. The chart contains the following rankings

- Blue represents the students who rated themselves with a 1
- Red represents the students who rated themselves with a 2
- Orange represents the students who rated themselves with a 3
- Green represents the students who rated themselves with a 4
- Purple represents the students who rated themselves with a 5

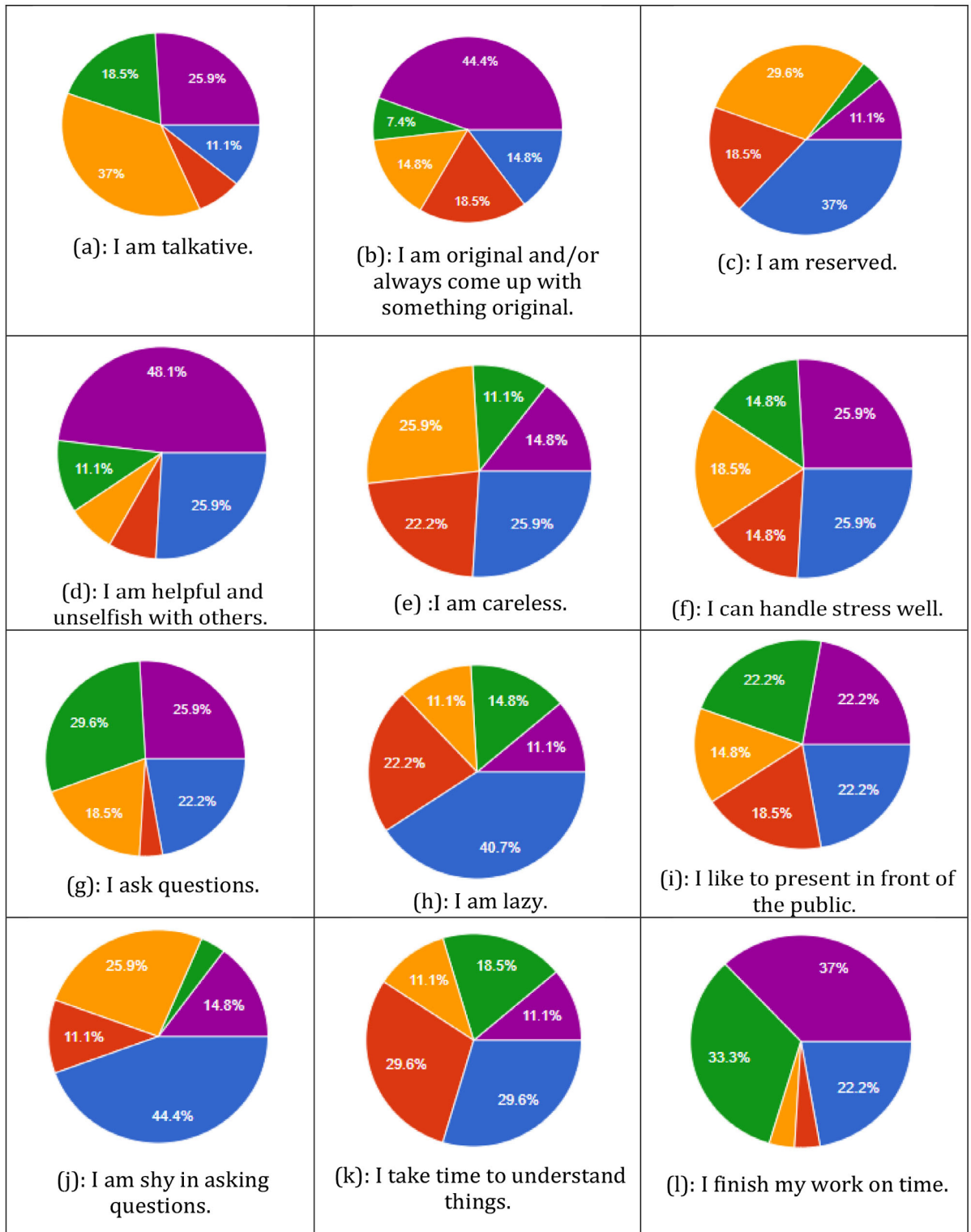


Figure 3: Results compiled by the proposed approach

B. Data Pre-processing

The second step is data preprocessing, in which a large amount of data is extracted and filtered. The randomly saved data is converted into usable and understandable sets of data. Then transformed it into structured data by storing it in the database. It is a very vital step to generate accurate results. Sometimes, a developer or researcher is tempted to skip this step and move to the next but clustering of unstructured data is not a good approach. When we have a big amount of data, a lot of data is not required and it should be filtered out. For preprocessing the data set, we perform the following steps:

- i. We prepared a google form and filled it by experts.
- ii. Data is extracted from google Forms.
- iii. Data cleaning is performed on extracted data like removing some extra columns.
- iv. The refine data has 27 rows and 26 columns
- v. K-mean clustering is applied to clean data
- vi. By applying the algorithm, we get the results

C. Data Clustering

Data Clustering is the third step. For categorizing, we applied k-means clustering to our structured data and, k as a result, we formed different sections of students. We set the value of k equal to the number of sections in which we want to divide our students, where k is equal to the total number of sections. We used the distance formula to determine the Euclidean distance as given in the following equation (i). The equation, 'd', represents the distance which can handle both ordinal and quantitative values. 'X' and 'Y' are the two coordinates or dimensions. A limit is applied on the dimensions, which is starts from '1' and runs till its value riches to 'p'

$$d(x,y) = \sum_{i=1}^p |x_i - y_i| \quad (i)$$

Pseudocode for k-means clustering Algorithm:

Inputs:

$D = \{t_1, t_2, \dots, t_n\}$ // student's data
K //total number of sections

Output:

K // k sections of students

//initialization

1. K-means algorithm:

2. Assign initial or random values for centroids (m_1, m_1, \dots, m_k)

3. For given values of iterations:

4. Iterate through random values:

5. Find the mean closest to the value

6. Assign a value to mean

7. Update mean

8. Repeat:

9. End

Assign the student each time to the section with which it has a minimum distance until convergence is achieved.

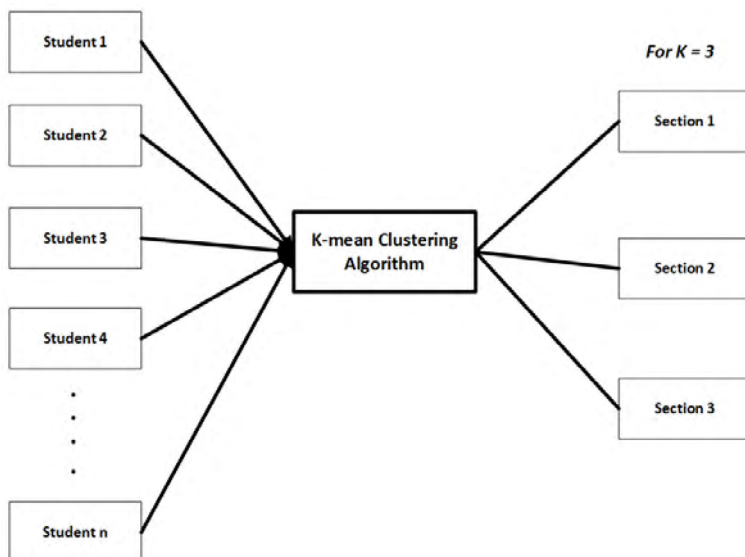


Figure 4: Demonstration of K-mean clustering algorithm for performing classification

D. Data Classification

The next step is data classification, in which the students with the same characteristics and habits are grouped and allotted the same section. Depending on the students' behavior, the most suitable teacher for them was assigned. This impact of the K-mean clustering algorithm is shown in Figure 4. This figure explains that there are a number of students. Every student has their own habits, characteristics, and behavior. When we applied the K-mean clustering algorithm to the students, it is fall in one section. Here we have a value of k is 3 so we have 3 sections, which means 3 types of categories for classification.

E. Data Storage

All the machine learning and artificial intelligence algorithms are based on data. Data is the lifeline in it and for the prediction of anything this data is used. The fifth and last stage is data Storage. All grouped data and information of students and teachers are stored in the database. Once data is collected, it must be stored in some reliable and secure database. The database is scalable thus that it can be increased or decreased when required.

4. Simulation and Results

The algorithm was applied and, as a result, regular subgroups or sections were made through the provided information. The algorithm divided the dataset into K pre-defined unique non-overlapping subgroups (clusters) where each information point has a place as it were one bunch. The following results were obtained after the first attributes were divided into 3 subgroups and the algorithm was applied.

In the first subgroup, the attributes considered were regarding the age and gender of the students, along with the attributes like students' analyzing behavior including talkativeness, innovation, reserved nature, carefulness, deep thinking skills, stress handling, and worrying. While considering all these attributes, k-means clustering (keeping $k = 3$) was applied as illustrated in Figure 5 (a). It divided students into 3 clusters (0,1,2) in the ratio [11:6:5]. The figure shows that those students having carefulness and deep thinking are less chance of stress and worry.

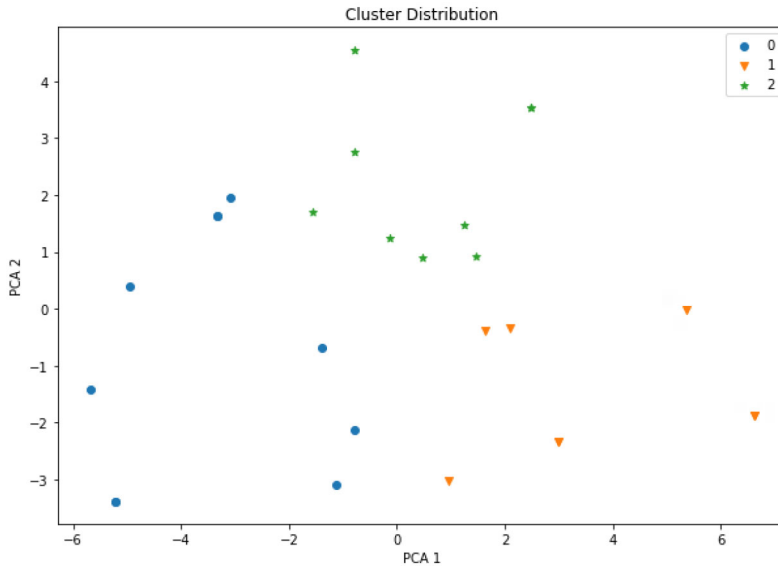


Figure 5 (a): k-means clustering (keeping k=3) on attribute set A

The second subgroup had students based on their age, gender and further attributes including laziness, confidence, shyness, obedience, punctuality, ability to work in groups, following directions while working, and understanding of nature. While considering the above-mentioned attributes, the K-means clustering algorithm (keeping k=3) was applied, as is illustrated in Figure 5 (b). This resulted in the division of students into 3 clusters (0,1,2) in the ratio [8:6:8]. When the graph reaches 0, as it gives the highest values apart from values that have medium or low points.

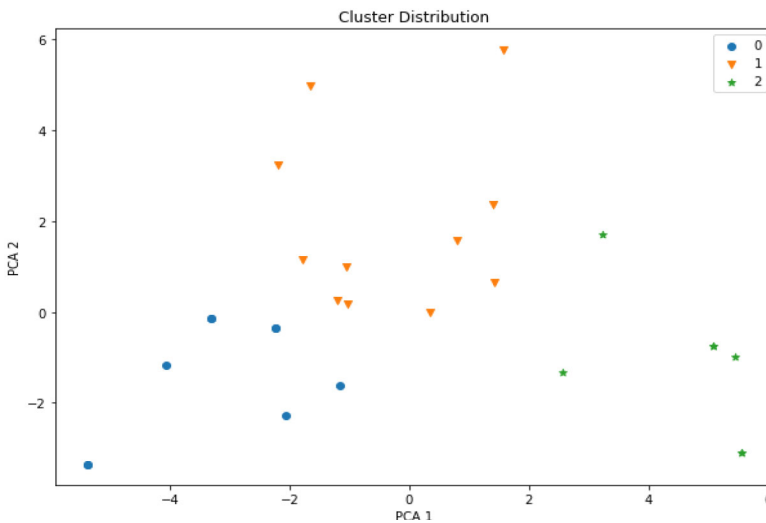


Figure 5 (b): k-means clustering (keeping k=3) on attribute set A

The attributes like age and gender of the students along with behaviors like deep thinking, stress handling, laziness, helping nature, worrying, interest in games and their frequent question asking were placed in the third subgroup. The k-means clustering algorithm (keeping $k = 3$) was applied by considering all of the above-mentioned attributes as illustrated in figure 5 (c). The students were divided into 3 clusters (0,1,2) in ratio [8:11:4]. Most of the values in the graph are high when they are away from the center value, which is zero. At zero, clusters have low values.

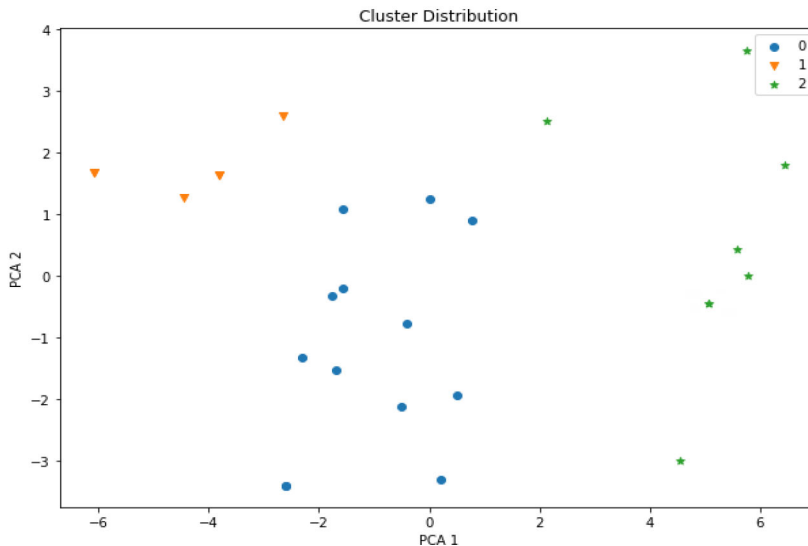


Figure 5 (c): k-means clustering (keeping $k=3$) on attribute set A

The algorithm was applied and all three sub-groups were merged after obtaining the above results. This was done while considering the attributes illustrated in Figure 5 (d). The division is made based on the students' having the same characteristics and behaviors. These are then placed in the same sub-group and there are a total of 3 sub-sections. The students were divided into 3 clusters (0,1,2) in the ratio [5:6:12]. Most values are running in a synchronized way, and the values have similarly high and low points.

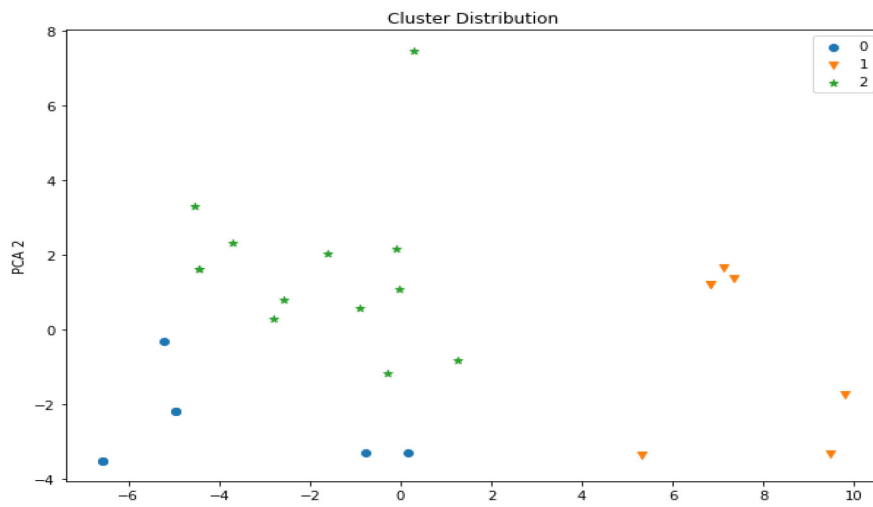


Figure 5 (d): K-means on all three combined

6. Conclusion

The previous studies and research carried out in the field of student behavior understanding and the required advancements were investigated and discussed. For the data collection, an online survey was designed and filled out by students up to class 5, as the proposed work mainly focuses on primary level children. K-means clustering algorithm was applied to the collected data and the sections of pupils were made based on similar characteristics, habits, and behavior. In the future, a new algorithm is being developed that will allow us to divide the student population into sections that are proportionally even as the current algorithm divides the children into an uneven proportion. Further literature review will be enhanced and experiments will be performed for the validation of our work.

Acknowledgment

This work is supported by PMAS-Arid Agriculture University funded project titled “Accurate sensing Information gathering by employing IoT in Agriculture of Pakistan” Project no. PMAS-AAUR/ ORIC/ 2026 Dated 08-03-2021.

References

- [1] F. Idris, Z. Hassan, A. Ya'acob, S. K. Gill, and N. A. M. Awal, "The Role of Education in Shaping Youth's National Identity," *Procedia - Soc. Behav. Sci.*, vol. 59, pp. 443–450, Oct. 2012, doi: 10.1016/j.sbspro.2012.09.299.
- [2] A. R. Monteiro, "The right of the child to education: What right to what education?," in *Procedia - Social and Behavioral Sciences*, Jan. 2010, vol. 9, pp. 1988–1992, doi: 10.1016/j.sbspro.2010.12.433.
- [3] V. K. Salgong, O. Ngumi, and K. Chege, "The Role of Guidance and Counseling in Enhancing Student Discipline in Secondary Schools in Koibatek District," *J. Educ. Pract.*, vol. 7, no. 13, pp. 142–151, 2016, Accessed: Jun. 21, 2021. [Online]. Available: <https://eric.ed.gov/?id=EJ1102862>.
- [4] K. Van Petegem, A. Aelterman, H. Van Keer, and Y. Rosseel, "The influence of student characteristics and interpersonal teacher behaviour in the classroom on student's wellbeing," *Soc. Indic. Res.*, vol. 85, no. 2, pp. 279–291, Nov. 2007, doi: 10.1007/s11205-007-9093-7.
- [5] R. K. Chaffee, A. M. Briesch, R. J. Volpe, A. H. Johnson, and L. Dudley, "Effects of a Class-Wide Positive Peer Reporting Intervention on Middle School Student Behavior:," <https://doi.org/10.1177/0198742919881112>, vol. 45, no. 4, pp. 224–237, Oct. 2019, doi: 10.1177/0198742919881112.
- [6] I. Albluwi, "Using Static Analysis Tools for Analyzing Student Behavior in an Introductory Programming Course," *Artic. Jordanian J. Comput. Inf. Technol.*, vol. 06, no. 03, 2020, doi: 10.5455/jjcit.71-1584234700.
- [7] E. Bernaras, J. Jaureguizar, and M. Garaigordobil, "Child and adolescent depression: A review of theories, evaluation instruments, prevention programs, and treatments," *Front. Psychol.*, vol. 10, no. MAR, 2019, doi: 10.3389/FPSYG.2019.00543/FULL.
- [8] J. Zong, C. Cui, Y. Ma, L. Yao, M. Chen, and Y. Yin, "Behavior-driven Student Performance Prediction with Tri-branch Convolutional Neural Network," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 2353–2356, Oct. 2020, doi: 10.1145/3340531.3412110.
- [9] R. K. Orr, P. Caldarella, B. D. Hansen, and H. P. Wills, "Managing Student Behavior in a Middle School Special Education Classroom Using CW-FIT Tier 1," *J. Behav. Educ.* 2019 291, vol. 29, no. 1, pp. 168–187, Apr. 2019, doi: 10.1007/S10864-019-09325-W.
- [10] W. Y. Jiao et al., "Behavioral and Emotional Disorders in Children during the COVID-19 Epidemic," *J. Pediatr.*, vol. 221, pp. 264–266, Jun. 2020, doi: 10.1016/j.jpeds.2020.03.013.

- [11] L. Chen, T. Q. Tang, Z. Song, H. J. Huang, and R. Y. Guo, "Child behavior during evacuation under non-emergency situations: Experimental and simulation results," *Simul. Model. Pract. Theory*, vol. 90, pp. 31–44, Jan. 2019, doi: 10.1016/j.simpat.2018.10.007.
- [12] R. Zheng, F. Jiang, and R. Shen, "Intelligent Student Behavior Analysis System for Real Classrooms," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2020, vol. 2020-May, pp. 9244–9248, doi: 10.1109/ICASSP40776.2020.9053457.
- [13] Z. Fan, J. Xu, W. Liu, and W. Cheng, "Gesture based misbehavior detection in online examination," in *ICCSE 2016 - 11th International Conference on Computer Science and Education*, Oct. 2016, pp. 234–238, doi: 10.1109/ICCSE.2016.7581586.
- [14] K. Godwin, V. Almeda, M. Petroccia, R. S. Baker, and A. V Fisher, "Classroom activities and off-task behavior in elementary school children," *Proc. Annu. Meet. Cogn. Sci. Soc. Title Classr. Act. off-task Behav. Elem. Sch. Child. Publ. Date 2013 Peer Rev.*, vol. 35, no. 35, pp. 2428–2433, 2013, Accessed: Jun. 21, 2021. [Online]. Available: <https://escholarship.org/content/qt8mx9h5hq/qt8mx9h5hq.pdf>.
- [15] I. Singh, "A disorder of anger and aggression: Children's perspectives on attention deficit/hyperactivity disorder in the UK," *Soc. Sci. Med.*, vol. 73, no. 6, pp. 889–896, Sep. 2011, doi: 10.1016/j.socscimed.2011.03.049.
- [16] S. Sarwar, "Influence of Parenting Style on Children's Behaviour," *J. Educ. Educ. Dev.*, vol. 3, no. 2, Dec. 2016, Accessed: Sep. 12, 2021. [Online]. Available: <https://papers.ssrn.com/abstract=2882540>.
- [17] R. E. Izzaty, "Happiness in early childhood," *Psychol. Res. Interv.*, vol. 1, no. 2, Sep. 2018, doi: 10.21831/pri.v1i2.22024.
- [18] N. N. R. Lima et al., "Childhood depression: a systematic review," *Neuropsychiatr. Dis. Treat.*, vol. 9, pp. 1417–25, Sep. 2013, doi: 10.2147/NDTS42402.
- [19] L. Safaei and M. Youzbashi, "Comparison of the Severity of Obsession and Working Memory in Children with Obsessive Compulsive Disorder and Healthy Children," *Int. J. Pediatr.*, vol. 8, no. 10, pp. 12275–12284, 2020, doi: 10.22038/ijp.2020.50337.4006.
- [20] U. Kessels and A. Heyder, "Not stupid, but lazy? Psychological benefits of disruptive classroom behavior from an attributional perspective," *Soc. Psychol. Educ.*, vol. 23, no. 3, pp. 583–613, Jul. 2020, doi: 10.1007/s11218-020-09550-6.
- [21] M. C. Levy, W. G. Kronenberger, and B. D. Carter, "Brief Report: Illness Factors and Child Behavior Before and During Pediatric Hospitalization," *J. Pediatr. Psychol.*, vol. 33, no. 8, pp. 905–909, Sep. 2008, doi: 10.1093/JPEPSY/JSN039.

- [22] A. D. Faigenbaum, L. A. Milliken, and W. L. Westcott, "Maximal Strength Testing in Healthy Children," *Natl. Strength Cond. Assoc. J. Strength Cond. Res.*, vol. 17, no. 1, pp. 162–166, 2003.
- [23] L. Mitchell, "NAUGHTY OR. NEEDED? EXCLUSIONS: A STUDY OF ONE LOCAL EDUCATION AUTHORITY," 1998. Accessed: Sep. 10, 2021. [Online]. Available: <https://etheses.whiterose.ac.uk/4243/1/DX206624.pdf>.
- [24] M. T. STEIN, A. GRAZIANO, B. HOWARD, and H. DUBOWITZ, "Maria : Stubborn, willful, and always full of energy," *J. Dev. Behav. Pediatr.*, vol. 17, no. 4, 1996.
- [25] T. L. Cross, "Nerds and Geeks: Society's Evolving Stereotypes of Our Students with Gifts and Talents:," <http://dx.doi.org/10.1177/107621750502800406>, vol. 28, no. 4, pp. 26–65, Jul. 2016, doi: 10.1177/107621750502800406.

Enhanced Accessibility of Facebook Messenger for Blind Users

Mamoona Atif Swati¹

Dr. Mustafa Madni²
Dr. Iftikhar Ahmed Khan⁴

Dr. Uzair Iqbal Janjua³

Abstract

With the growth in technology, social networking has become an essential factor in human life. People connect and share information through social media applications like Instagram, Facebook, and Twitter. Though, it is witnessed that using such applications is challenging for blind users. Such applications are also stated to be incredibly inaccessible. This study examines the usefulness of the Facebook Messenger application by using smartphone devices for visually impaired and blind users. Firstly, a pilot experiment is conducted with five selected blind people, and their performance and interaction are observed with the existing Facebook Messenger application. A prototype is designed and implemented based on existing Web Content Accessibility Guidelines (WCAG) to minimize the difficulties observed during the initial experiment.

Further, twenty-one blind users experienced the proposed prototype and the existing messenger application. The findings have shown that the proposed prototype design fulfills the efficiency and user satisfaction for blind users. Finally, future work is recommended based on the acquired outcome to enhance the usability of social media applications.

Keyword: Accessibility, social networking, blind users, WCAG 2.0 guidelines.

1. Introduction

In recent years, smartphones have increased and will reach 6,648 million users by 2022[1]. Smartphones are generally used to communicate, share photos and videos, play games, use social media applications, etc. Social media applications play a vital role in sharing information and a central role in socializing with people. It is stated that compared to other media platforms, Facebook is 67.4% popular [2]. Therefore, such applications must be adaptive to every environment and accessible to most users. However, blind or Visually Impaired (VI) people cannot utilize the applications effectively and efficiently because of accessibility issues, especially while accessing graphics [24]. Accessibility allows blind users to access the features virtually [3]. Facebook Messenger is generally accepted and widely considered more convenient than the Facebook website [16].

¹Department of Computer Science | COMSATS University Islamabad | mamoonaswati393@gmail.com

²Department of Computer Science | COMSATS University Islamabad | tahir_mustafa@comsats.edu.pk

³Department of Computer Science | COMSATS University Islamabad | uzair_iqbal@comsats.edu.pk

⁴Department of Computer Science | COMSATS University Islamabad, Abbottabad | iftikharahmed@cuiatd.edu.pk

However, Facebook Messenger applications should be developed under a comprehensive perspective to provide equal access to every citizen. Hence, the effectiveness of social networking applications is evaluated with people without any visual impairment. Though, the concerns are more significant for the part of VI people. As reported by World Health Organization, there are 285 million VI, out of which 39 million people are fully blind [4]. Hence, using mobile or social media applications is challenging with such impairment. Thus, blind people need a sighted person for assistance to access an application, which is helpful, but it is not prudent as sighted people are not always around or available.

Moreover, it is stated that 75% of people think that highly accessible applications are well-developed with assistive technology [5]. Above all, the complexity of the application is another challenge in accomplishing various tasks. The social media applications like Facebook Messenger have complex tasks, for example, creating a group, sending a message, deleting a conversation, etc.

There are several proposed guidelines to make the applications accessible and usable for blind users, like the World Wide Web Consortium's (W3C) Web Content Accessibility Guidelines (WCAG) [6]. However, WCAG 2.0 has given the means to measure the website's accessibility. Even though there are no such well-recognized guidelines to assess the accessibility of mobile phone applications [7]. The absence of implementing such procedures in websites and applications is also listed in the literature as a significant issue [10] - [15] that reduces the accessibility and loss of control. Therefore, it generates uncertainty in understanding the information, which reduces the interest of blind people and affects their overall performance [14]. Hence, examining the interaction between social media and blind users is necessary. This study aims to develop an accessible prototype of a messenger application based on problems identified in previous work [23]. Moreover, an experiment is performed to measure the performance of the proposed messenger using the System Usability Scale SUS questionnaire [9].

The rest of the paper is formatted as follows: Section two is the literature review that discusses the state-of-the-art techniques in social media and accessibility. Section three describes the proposed methodology, followed by section four, which presents our results. The last sections contain the conclusion and future work.

2. Literature Review

Social networking sites are a communication bridge between people. People communicate and share information on social media [17]. However, due to some accessibility issues, VI or blind users cannot access social networking sites correctly. Also, blind people conceptualize web interaction differently than sighted people [18].

Some solutions are already provided for the difficulties faced by blind users. World Wide Web Consortium provided design standards for developers to make a web application accessible [3]. Furthermore, authors have developed accessible smartphone applications for blind users. The author created a flexible wayfinding smartphone application for all users. The application was further tested by eight VI users, which showed that VI users could use the application without any help [19]. In another study, VI and blind users experienced visual content on social networking sites and conducted qualitative research to discover the challenges, practices, and experiences of blind users [8].

Moreover, the author identified the use of social media in rural and peri-urban India [20]. They have also observed how the participants used computers, social media platforms, and smartphones. Further, the weaknesses and strengths of Facebook, WhatsApp, and Twitter for blind people are discussed, and a thorough analysis of how blind people with less income in India have adopted a social media voice opportunity.

Another study identified problems with the Facebook homepage [21]. The Facebook homepage interface was redesigned to an accessible version using HTML 5 and Human-Computer Interaction guidelines and was further evaluated. Furthermore, the author assessed the affects caused in VI people while using the features of Facebook and compared their experience to the experiences of sighted users. Once the author collected the information, statistical analysis was performed to estimate users' feelings [22].

After studying various research contributions, it was evident that VI and blind people face inaccessibility issues, particularly with Facebook. Also, it is observed that there is little or almost no work done on measuring Facebook Messenger's accessibility. Moreover, the majority of the work is based on the website.

3. Proposed Methodology

Most of the present literature on Human-Computer Interaction (HCI) lies in the paradigm of positivism with the critical concept of discovering the undiscovered. Positivism falls toward the quantitative dominant research approach as an empirical phenomenon is considered to yield empirical prediction. In this study, as shown in figure 1, a quantitative dominant research approach has been followed using a controlled experiment method to define the real cause of the phenomenon and fundamental casual relations.

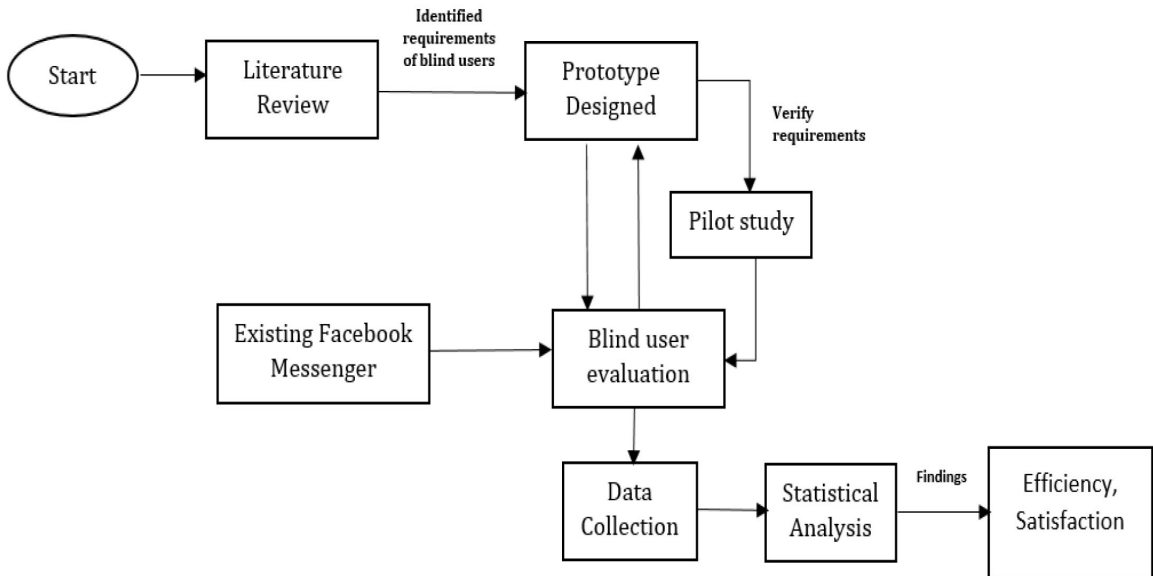


Figure 1: Block Diagram of the proposed methodology

The study is performed in two phases. Initially, the user's opinion is assessed for the existing Facebook Messenger, which identifies the user's thought through the tasks listed in Table I and the importance of WCAG guidelines through the current Facebook messenger [23]. Although, in this research, a comparative evaluation is performed among the existing Facebook messenger and the proposed messenger. During the experiment, the participant needed to narrate the tasks verbally while performing. The participants started describing the study, but it was observed that it slowed their productivity, and the participants' attention was distracted while explaining their job while interacting, which resulted in making mistakes.

Table 1: Tasks performed by blind users

No.	Tasks
T1	Create a new group
T2	Add participants to the group
T3	Send a message in the group
T4	Create an admin
T5	Delete Conversation
T6	Modify the group name
T7	Send a voice message and send it to the group
T8	Leave the group

A. PHASE II: Evaluation of Proposed Facebook Messenger

This section gives the quantitative approach of the experimental studies based on interaction experiences of both the proposed and existing messenger using a multi-touch smartphone device.

i. Goals

The study's primary goal was to evaluate the usability of the two different messengers, i.e., existing Facebook Messenger and the proposed messenger for social networking for blind users using smartphones. This study may be beneficial in determining the usefulness of the messenger application for social networking from the user's perspective.

ii. Device Used

The device used is the Samsung Galaxy J6 smartphone with Android Operating System, a commercially available touch-screen phone with improved image quality. This system is equipped with the Tru-Octa Core processor and 3.0 GB RAM. The Messenger applications are adequately adjusted before experimenting.

iii. Participants

Twenty-one participants (VI and blind users) were involved in this study. All the participants were chosen from the Government school of blind, Shamsabad in Rawalpindi, Pakistan. Amongst these participants, 11 were males, and 10 participants were females. The age of participants was between 18 to 36 years old. The selection criteria are based on their skills and experience using smartphones, familiarity with assistive technologies, and using existing Facebook Messenger.

iv. Task

For this study, the tasks were selected based on sharing information between group members, interaction, and transmission using social media applications. The jobs are: create a group, add participants in-group, create admin, modify group name, delete a conversation, and leave a group.

v. Hypothesis

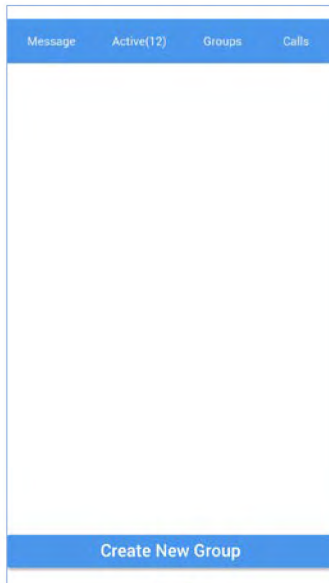
A null hypothesis is formalized to determine the performance of the system. It is stated as there is no difference in task completion time when using the existing Facebook Messenger and the proposed messenger.

vi. System Development

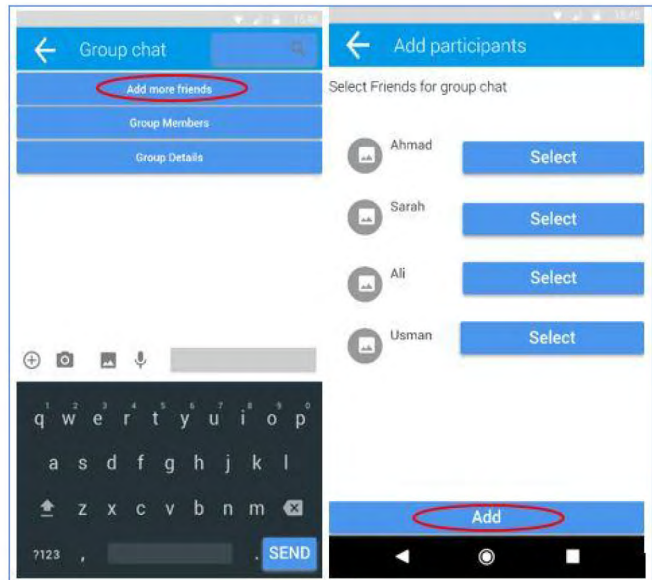
The proposed messenger was designed and developed as an interactive and functional high-fidelity prototype. These prototypes were created using the Justin Mind tool. It allows the development of an interactive mobile application for the selected tasks. For each job, audio feedback is provided to the user to minimize confusion, and users do not need any assistance from other people.

The first task was to create a group. The button to create a new group is located at the end, as shown in Figure 2 (i). When the switch is tapped, it gives audio feedback as 'create a new group.' Once the group has been created, it provides feedback to the user as 'group created successfully. Once the group is completed successfully, the user can add a participant in-group and start a group chat. The audio feedback facilitates the user, and s/he does not need assistance from others. It differentiates between adding participants while creating a group and adding more friends after the group is completed, as shown in Figure 2 (ii).

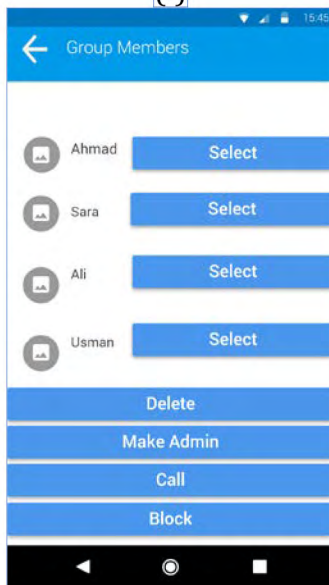
Users can select any member and can create an admin. Furthermore, users can delete, block, or call any selected member, as shown in Figure 2 (iii). Moreover, the group details are categorized in the group details. Also, appropriate menu categories are adopted, so the user can easily access the digital content, as shown in Figure 2 (iv). The group details include the properties of a group such as a share link, request a member, delete a conversation, change color, emoji, name or photo, etc.



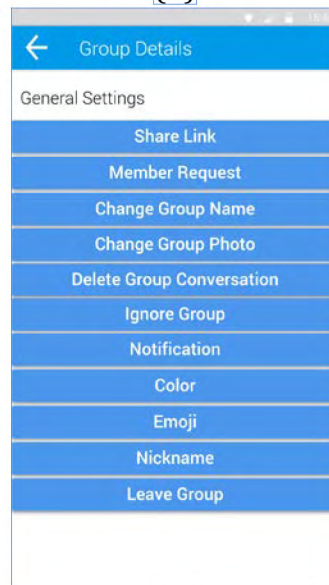
(i)



(ii)



(iii)



(iv)

Figure 2: UI design of proposed messenger (i) create a new group, (ii) Add participant to a group, (iii) group members, and (iv) group settings

In this development of the proposed messenger, the aim was simplicity and minimized navigation depth compared to the existing Facebook messenger. Hence, the depth of the tasks (such as changing the group name, deleting the conversation, and leaving the

group) is minimized to increase the simplicity and reduce the navigation steps. As a result, blind people will efficiently perform all tasks in fewer steps. There are five steps in the existing Facebook messenger to change the group name, whereas, in the proposed messenger for the blind, the user only needs three simple steps to complete the same task, as shown in Figure 3 (i). The user needs five steps in existing Facebook Messenger to delete a conversation. As depicted in Figure 3 (ii), the user can complete the task in three simple steps in the proposed messenger. The navigation steps to reach an icon are reduced and minimized. Similarly, the user needs five efforts to leave a group in the existing Facebook messenger. Users only need three steps to complete the same task in the proposed messenger, as shown in Figure 3 (iii).

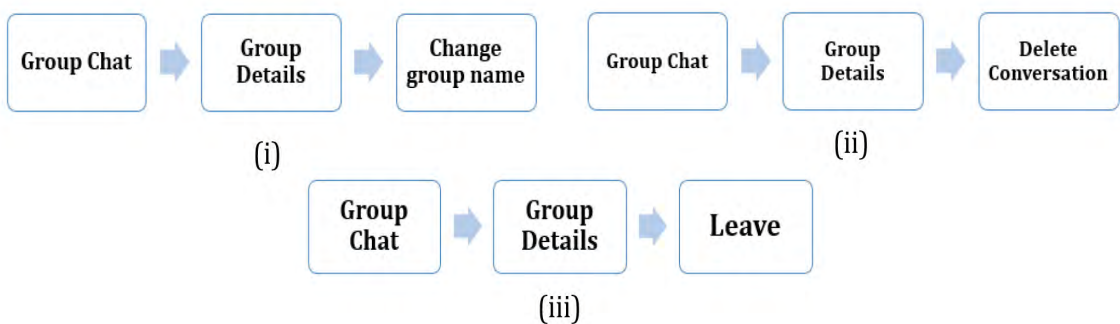


Figure 3: Steps while performing tasks (i) change group name, (ii) delete a conversation, and (iii) leave the group.

i. Procedure

The setting and smartphone setup for this study were organized in the laboratory in the Government school of blind, Shamsabad in Rawalpindi, Pakistan. The laboratory provides a comfortable seating arrangement for test participants. The procedure of this study is divided into three stages: before the experiment, i-e, the initial stage, to get consent and fill in demographic information, one which comprises the training and assessment of several tasks, and finally gathering the feedback of participants.

In the initial stage of the experiment, consent forms were read aloud to all the participants. The participants agreed to the terms and allowed us to record them. Afterward, they were repeatedly told the time to start the experiment and were provided time to fill in demographic information.

In the second stage, all the participants were given a comfortable seating arrangement, and all tasks were explained again to clarify their doubts about the participants. Afterward, the participants randomly experienced existing Facebook messenger and proposed messenger using a smartphone.

In the third stage, after the participants randomly experienced both the proposed and existing messenger, they presented their feedback using objective measure (efficiency) and subjective measure (user satisfaction by System Usability Scale questionnaire).

4. Experimental Analysis

The experiments are based on discovering the Facebook messenger's accessibility for blind users. Hence, phase I of the study explores the user's perspective about the tasks and shows the importance of Web Content Accessibility guidelines through existing Facebook messenger. Though, it is noticed that blind users face various issues while using such applications. According to the identified findings, the succeeding experiment is implemented, empirically assessing the accessibility of both existing and proposed Facebook messenger. The results of Phase 1 are mentioned in the previous research paper [23].

The previous study showed that blind users faced problems while performing the tasks, but they could finish them with human assistance. The responses of the blind people were compared with the WCAG guidelines, as shown in Table II, and discussed in the prior study [23]. These guidelines have four principles, i.e., perceivable, robust, operable, and understandable[6].

Table 2: Comparison of problems with WCAG guidelines

Principles	Description	Task Identified with problem related to principle
Perceivable	The user interface (UI) must provide the exact message to the user that is intended for the user.	T1, T2, T5 and T7
Operate-able	All the components and the UI must be operate-able by users.	T1, T2, T4, T6, and T9
Understandable	The operation and the content on the UI are easily understandable for the users.	T1, T2, T3, T4 and T8
Robust	The content on the UI should be perceived reliable by most of the users and the assistive technologies e.g. screen reader for VI	None

This section illustrates the findings of the research on both interfaces. It empirically assesses the accessibility of both the existing Facebook messenger and the proposed messenger. The data were converted from the questionnaire and objective measurements to the SPSS tool for statistical analysis. In total, 3 out of 24 participants stopped due to the incapability of using a smartphone. Of these 21 participants, 10 were females, and

11 were male. The gathered data is cleaned before applying any statistical analysis by removing missing values and outliers using a box plot.

Initially, efficiency is measured to complete all tasks of the application. Most participants spent time assigning an admin and changing the group name. Figure 4 shows the time calculated to achieve the existing Facebook messenger and proposed messenger tasks.

An independent t-test was conducted to compare the mean for existing and proposed Facebook messenger. Every participant performed all eight tasks on both interfaces. The time for every job is depicted in Table III. The result of the Independent T-Test showed a significant difference in the scores for existing Facebook messenger ($M=21.3$, $SD=3.59$) and proposed Facebook messenger ($M=9.7$, $SD= 2.2$) conditions; $t(40) = 12.54$, $p= 0.00$; hence H_0 is accepted. Overall results of the T-Test propose that the time to complete tasks in the proposed messenger is less than the existing Facebook messenger.

Table 3: Task Completion Time of Existing Facebook Messenger and Proposed Messenger

Tasks	Existing Facebook messenger (Mean \pm SD)	Proposed messenger (Mean \pm SD)
T1: Create a new group	19.85 \pm 7.185	14.54 \pm 6.99
T2: Add participants to the group	15.48 \pm 4.14	7.98 \pm 5.05
T3: Send a message in the group	26.78 \pm 6.76	3.29 \pm 1.99
T4: Create an admin	51.15 \pm 2.68	19.58 \pm 8.67
T5: Delete Conversation	9.97 \pm 14.04	6.64 \pm 5.76
T6: Modify the group name	32.89 \pm 13.92	12.72 \pm 7.19
T7: Send a voice message and send it to the group	20.19 \pm 1.93	6.19 \pm 1.73
T8: Leave the group	5.13 \pm 2.09	5.96 \pm 3.9

The efficiency of both interfaces is compared by comparing the mean values, as depicted in figure 8. Hence, the proposed messenger consumes less time to finish tasks than the existing Facebook messenger application. The chart shows that the time consumed to create an admin and send a message is more significant in the current Facebook Messenger than in the proposed messenger due to usability issues and lack of feedback.

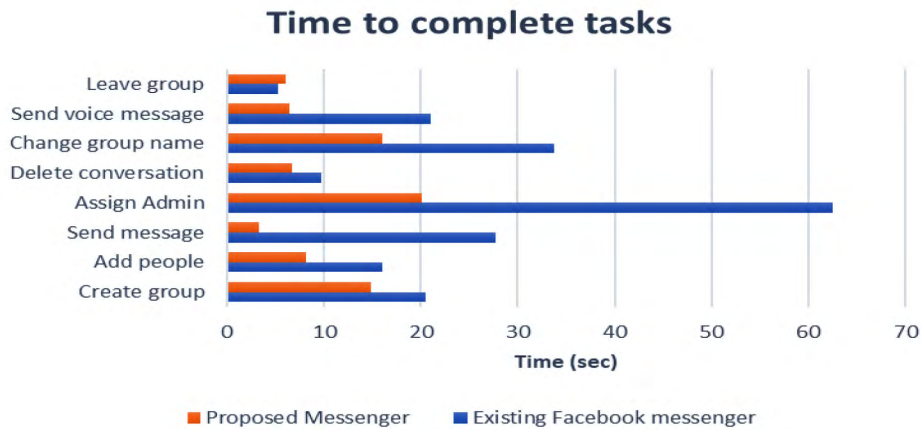


Figure 4: Time to complete tasks in both interfaces

An interview has been conducted to measure the usability of the proposed interface of Facebook messenger for blind users. The obtained data have a Cronbach's Alpha value of 0.73. To analyze the SUS questionnaire, the scale of odd questions is subtracted by one, and the scale of actual questions is removed by 5. The final result is added and further multiplied by 2.5 [26]. A result below 70 means that the system has some usability issues, and scores above 70 are better [27]. The results are above average.

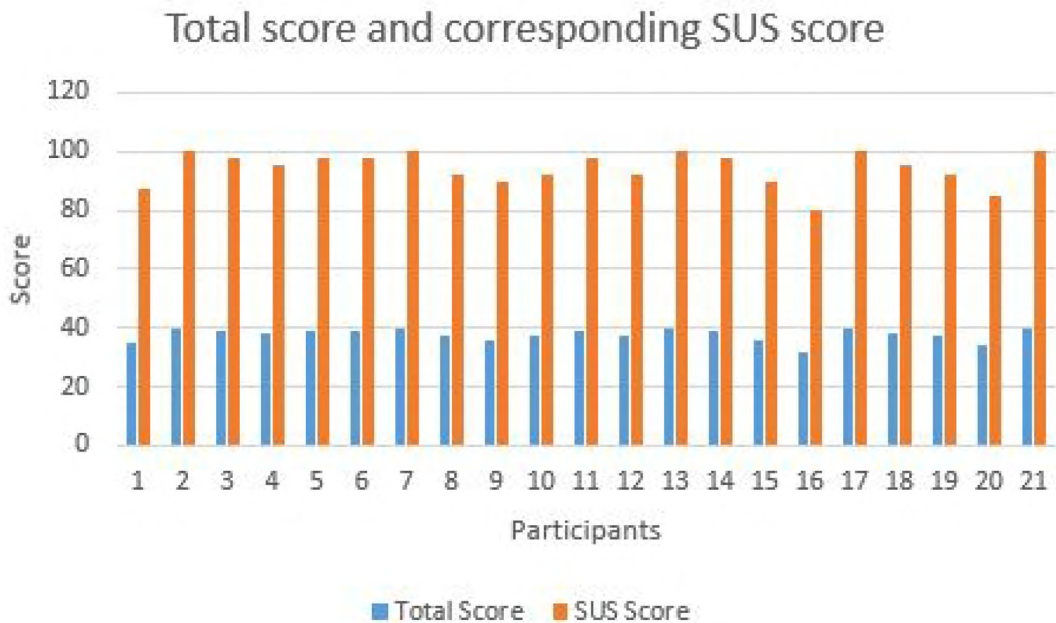


Figure 5: Total score and corresponding SUS Score

The graph presented in Figure 5 shows the overall and corresponding SUS scores. The Cronbach Alpha Test is applied to the questionnaire to measure the consistency between the related items. According to the rule of thumb, the reliability coefficient is only acceptable if the result is above .70 or higher [27]. The alpha coefficient for 21 items of our questionnaire is 0.733, suggesting that the items have high internal consistency.

5. Discussion

The previous study's findings exposed that blind users face problems accessing social media because of the inaccessible content [23]. Therefore, several guidelines are suggested to make the mobile applications and websites more accessible for VI and blind users, such as Web Content Accessibility Guidelines (WCAG) and World Wide Web Content (W3C). The absence of such guidelines results in the lack of accessibility to social networking applications. Subsequently, the interest of blind users and overall performance is affected by the absence of accessibility to social media applications.

The study highlighted that the number of Facebook users worldwide is 67.4%, a comparatively higher value [2]. Thus, it indicates that Facebook is used more than other social media mobile applications. However, Facebook is inaccessible for blind users because of its great visual content. As a result, an overall interview was conducted with blind users related to the Facebook Messenger mobile application. Hence, it is discovered that blind users and VI have not yet identified group chat features in Facebook Messenger. Therefore, Facebook group chat in Facebook messenger aids fascinating social options for all blind users, such as updating the group name, creating a group for people with the same interests, sending a message to various people at a time, and deleting the entire group chat at once, etc. Thus, formative experiment research was conducted (Phase I) to present this feature and identify accessibility problems that blind users encounter.

Moreover, the experiment was qualitative research-based. As we recognized through an investigation in Phase I, blind users faced issues related to navigation, understanding the context, and accessing the menu items [23]. Hence, the existing guidelines can be used to fix such identified challenges.

Furthermore, a summative experiment study was conducted to measure efficiency and satisfaction. Hence, to calculate the efficiency and identify a highly accessible interface for the blind, an experiment was conducted randomly on both interfaces by the same participants. One experiment is performed on the available Facebook messenger application, and the second experiment is performed on the Facebook Messenger Prototype for blind people. Each participant is supposed to serve all similar tasks on both interfaces. We observed and noted the time consumed for the participant to complete every task.

A. *Experiment on Existing Facebook Messenger*

The outcomes showed all 21 participants completed the first task, while some of the participants consumed time to finish the task as they could not find create a new group button. Participants 2, 17, and 19 stated that the excessive amount of options makes it difficult for them to find the button. The participants did not face difficulty adding people and finished the task in less time except for participant 2. The participant needed guidance in finding the button to add friends. While sending a message, all the participants effectively completed the task, but participant 6 faced a little trouble finding the send button.

The participants successfully reached the voice message icon but with trouble. However, making a member admin of the group was time-consuming and challenging. As mentioned earlier, the task includes six navigation steps, which increases the time to finish the task. The participants were able to change the group name, but few of them consumed time. Once the participants reached the button, they could quickly delete the conversation and leave the group in less time.

B. *Experiment on Advanced Facebook Messenger Prototype*

The results determine that the experiment performed on the Facebook Messenger prototype consumed less time finishing the tasks than the already available application. As a result, participants did not face any difficulty or confusion in completing the tasks. The reason is appropriate audio feedback and applied guidelines on the prototype. The audio feedback helped participants in navigation. As for sending a voice message, the voice icon explains how to send a voice message: hold the button, record the audio, and then release it; once the participants release the controller, the message will be sent to the group. Hence, the time to reach the button is observed and noted, while one of the tasks that took time was to assign an admin. The outcome illuminates that every participant completed all tasks in less time, except Participant 11 took the time to finish the first task. Participant 2 consumed time to change the group name. The result of the experiment proved that the proposed Facebook Messenger prototype is accessible to blind users compared to the Available Facebook Messenger.

This paper aims to resolve the issues related to the accessibility of the Facebook Messenger mobile application. A contribution was added to the discussion by offering blind people's experiences with selected tasks in the messenger. The meeting was held around the challenges faced by blind people and solutions provided for them.

6. Conclusion

We recognized the issues and difficulties which were faced by the participants while completing the tasks. Then their identified problems were noted down. Most of the issues encountered by the participants were related to the items, understanding the main context of the task and content, recognizing the menu, and navigation. Additionally, these problems were compared with the four WCAG 2.0 principles, i.e., perceivable, understandable, robust, and operable.

We recognized the issues and difficulties faced by the participants while completing the tasks. Then their identified problems were noted down. Most of the issues encountered by the participants were related to the items, understanding the main context of the task and content, recognizing the menu, and navigation. Additionally, these problems were compared with the four WCAG 2.0 principles, i.e., perceivable, understandable, robust, and operable.

WCAG 2.0 guidelines were recommended and implemented to reduce the problems and difficulties for the blind in accessing social media applications. There were 12 recommended guidelines in the previous paper to increase the accessibility of the Facebook messenger application. In this paper, we have developed a Facebook messenger prototype to solve the accessibility problem of the Facebook messenger for blind people. A final controlled experiment is conducted to measure the time and satisfaction between both interfaces. Furthermore, time is measured through a t-test. The result of the t-test rejects the null hypothesis and proves that both systems are not equal.

Our research has identified several essential details. We can suggest the following research directions: The study is open to identifying accessibility problems faced by VI and blind users while accessing other social media applications, e.g., Instagram, Snapchat, and WhatsApp. A set of guidelines can be compared with the problems identified by blind users. With the help of these guidelines, we can implement the interfaces of other social media applications specifically for Blind users. Moreover, the satisfaction and usability of implemented interfaces can be further compared with existing interfaces of other social media applications. Besides, this experiment is conducted for a particular area and in a specific region. The data can be further collected from all the country's regions to make it a generic model.

References

- [1] Statista, "Smartphone users worldwide." [Online]. Available: <https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/>. [Accessed: 25-Feb-2021].
- [2] GlobalStats, "Social Media Stats." [Online]. Available: <http://gs.statcounter.com/social-media-stats/all/pakistan#monthly-201707-201807>.
- [3] R. Babu, (2011) "Developing an Understanding of the Nature of Accessibility and Usability Problems Blind Students Face in Web-Enhanced Instruction Environments,," ISBN: ISBN-978-1-1248-7865-2.
- [4] W. H. ORGANIZATION, (2010) "Global Data on," p. 17.
- [5] A. M. Michalska, C. X. You, A. M. Nicolini, V. J. Ippolito, and W. Fink, (2014) "Accessible Web Page Design for the Visually Impaired: A Case Study," *Int. J. Hum. Comput. Interact.*, vol. 30, no. 12, pp. 995–1002.
- [6] M. Cooper and L. G. Reid, (2014) "Web Content Accessibility Guidelines (WCAG) 2.0," pp. 1–33.
- [7] Park, K., Goh, T., & So, H. (2014). Toward accessible mobile application design: developing mobile application accessibility guidelines for people with visual impairment.
- [8] V. Voykinska, C. Tech, and G. Leshed, (2016) "How Blind People Interact with Visual Content on Social Networking Services," DOI:10.1145/2818048.2820013. pp. 1584–1595.
- [9] Y. Borodin, J. P. Bigham, G. Dausch, and I. V Ramakrishnan, (2010) "More than meets the eye: A Survey of Screen-Reader Browsing Strategies," *Proc. 2010 Int. Cross Discip. Conf. Web Access. - W4A '10*, pp. 1–10, 2010.
- [10] C. N. Nobre, M. R. G. Meireles, and D. B. F. D. A. Silva, (2018) "Emotionally Oriented Analysis of the Experiences of Visually Impaired People on Facebook," *Journal: ACM Transactions on Accessible Computing*, vol. 11, no. 3.
- [11] W. Seo and H. Jung, (2018) "Understanding Blind or Visually Impaired People on YouTube through Qualitative Analysis of Videos," *Proc. 2018 ACM Int. Conf. Interact. Exp. TV Online Video - TVX '18*, pp. 191–196.
- [12] A. Vashistha, E. Cutrell, N. Dell, and R. Anderson, (2015) "Social Media Platforms for Low-Income Blind People in India," *Proc. 17th Int. ACM SIGACCESS Conf. Comput. Access. - ASSETS '15*, pp. 259–272.
- [13] E. Brady, M. R. Morris, and J. P. Bigham, (2015) "Gauging Receptiveness to Social Microvolunteering," *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. - CHI '15*, pp. 1055–1064.

- [14] S. Wu, J. Wieland, O. Farivar, and J. Schiller, (2017) "Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service," ACM Conference: Cscw, pp. 1180–1192.
- [15] J. A. Smith, M. R. Lind, D. Colton, D. Colton, and S. Hunsinger, (2010) "Website Accessibility for Users with Visual Impairment," Journal: Information System Education Journal, vol. 8, no. 53.
- [16] Voykinska, Violeta & Azenkot, Shiri & Wu, Shaomei & Leshed, Gilly. (2016). How Blind People Interact with Visual Content on Social Networking Services. DOI:10.1145/2818048.2820013.
- [17] De Nooy, W., Mrvar, A., & Batagelj, V. (2018). Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software (3rd ed., Structural Analysis in the Social Sciences). Cambridge: Cambridge University Press. doi:10.1017/9781108565691
- [18] R. Babu, (2014) "Can Blind People Use Social Media Effectively? A Qualitative Field Study of Facebook Usability," Am. J. Inf. Syst., vol. 2, no. 2, pp. 33–41.
- [19] M. C. Rodriguez-Sanchez, M. A. Moreno-Alvarez, E. Martin, S. Borromeo, and J. A. Hernandez-Tamames, (2014) "Accessible smartphones for blind users: A case study for a wayfinding system," Expert Syst. Appl., vol. 41, no. 16, pp. 7210–7222.
- [20] A. Vashistha, E. Cutrell, N. Dell, and R. Anderson, (2015) "Social Media Platforms for Low-Income Blind People in India," Proc. 17th Int. ACM SIGACCESS Conf. Comput. Access. - ASSETS '15, pp. 259–272.
- [21] P. Grober and J. Koster, (2017) "An Analysis to Overcome Shortcomings to Improve the Accessibility for the Blind: A Case Study on Facebook's Homepage," Proc. - 12th Int. Conf. Signal Image Technol. Internet-Based Syst. SITIS 2016, pp. 442–449.
- [22] C. N. Nobre, M. R. G. Meireles, and D. B. F. D. A. Silva, (2018) "Emotionally Oriented Analysis of the Experiences of Visually Impaired People on Facebook," Journal: ACM Transactions on Accessible Computing, vol. 11, no. 3.
- [23] Mamoona Atif Swati, Tahir Mustafa Madni, Uzair Iqbal janjua and Iftikhar Ahmad,(2021) " Accessibility of Social Media Application for Blinds," Conference of ICCIS
- [24] S. Reinders, (2021), " Accessible interactive 3D models for blind and low-vision people", ACM SIGACCESS Accessibility and Computing, issue 129, Article No. 6, pp 1-7

Automatic Speech Recognition on Non-Pathological Dataset of Urdu Language

Anoshia Imtiaz¹
Hira Zahid⁴

Munaf Rashid²
Muzzaffar Iqbal⁵

Sidra Abid Syed^{3*}
Akhtar Ali Khan⁶

Abstract

Voice is a primary tool for communications and voice disorders bring atypical characteristics in the voice which influence the quality of voice. Voice disorders are abnormal conditions that influence the quality of voice. Several protocols, including acoustic analysis, can detect clinical voice pathology. Based on the computerized acoustic analysis, machine learning algorithms and non-invasive systems may play a vital part in the initial detection, tracking, and growth of proficient pathological speech analysis. The methodology proposes to collect a non-pathological dataset, i.e., a healthy voice dataset, and offers a unique combination of feature extraction techniques combining Mel-Frequency Cepstral Coefficients (MFCC), and Pitch. Support Vector Machine (SVM) was used as a machine learning classifier for the training and testing of the dataset model. The SVM algorithms demonstrated satisfactory training and testing accuracy rate, i.e., 85.886%, which proves to be a milestone on the Urdu language dataset.

Keyword: Voice dataset, Urdu language, SVM, MFCC, Pitch.

1. Introduction

The human voice is the fundamental means of communicating and delivering verbal meaning [1]. It has been reported by the ASHA (American Speech–Language–Hearing Association) [2] that vocal disorders, also known as voice pathology, have a significant influence on people's day-to-day and professional lives. Disordered voices can cause social disadvantages and inferiority complexes, especially those already marginalized. When a person's quality, pitch, or volume differs or is unsuitable for their age, sex, culture, or geographic region, they typically report having a voice problem. Expresses concern over

^{1,4,5,6}Biomedical Engineering Department, Ziauddin University Faculty of Engineering Science Technology and Management, Karachi, Pakistan

²Electrical Engineering Department & Software Engineering Department, Ziauddin University Faculty of Engineering Science Technology and Management, Karachi, Pakistan

³Biomedical Engineering Department, Sir Syed University of Engineering & Technology, Karachi, Pakistan
Corresponding Author email id: sidra.gha@yahoo.com

developing a voice that differs from one's usual one and does not satisfy basic demands, although others do not notice the difference [3]. Generally, sounds may be divided into three categories: voiced sounds (such as vowels and nasals), unvoiced sounds (such as fricatives), and stop-consonants (e.g., plosives). Although speech originates in the lungs, it is created when air travels through the larynx and vocal cords [4]. The sound can be categorized as follows based on the health of the larynx's vocal folds: the time-periodic and harmonic voiced sound; the more noise-like unvoiced sound [5]. Speech processing is a broad and well-studied issue with applicability in telecommunications, audiovisual, and other disciplines. Real-time speech processing offers more problems than offline action. The processing includes various features, such as differentiating between utterances, identifying the speaker, etc.

Waveform and source coding are the two central coding schemes used in speech modeling [6]. When the researchers first started, they tried to copy the sounds exactly, which they dubbed waveform coding. This technique uses quantization and redundancies to try to keep the actual waveform. Instead of separating the sound into different components, it can be divided up and then each one can be modeled independently. Source coding is the term used to describe this approach of employing variables. For the identification of the spoken phrases, gender, identification of the speaker, and further speech aspects might be used. Pitch is one of the essential aspects of speech. The difference in pitch between speech signals is substantial. Vocal fold oscillation frequency influences pitch: for example, a rise of 300 Hz is produced by oscillating the folds 300 times per second. While the air travels through the folds, integer iterations of the fundamental frequency (harmonics) are also made—the pitch changes with the singer's age. In the period leading up to adulthood, the pitch is about 250 Hz. Slope ranges from 60 to 120 Hz for mature males and 120 to 200 Hz for adult women [7]. For generating speech, Rabiner, and Schafer's [8] discrete-time model utilizes linear prediction. An impulsive oscillator simulates voiced speech excitement; a glottal shaping filter then processes the impulses. A random noise generator generates the unvoiced speech. Ideally, any characteristics chosen for a speech model (1) should not be purposefully influenced by the speaker, (2) should not be independent of their physical state, and (3) resistant to any ambient noises. Although a speaker's pitch may be readily adjusted, it can also be a low-pass filter that can be applied as a feature to remove any noise and interference.

Much study has been conducted on voice processing/recognition in English, Spanish, German, and Arabic, but implementing these time-tested methods to the Urdu language has not been explored much. As a result, we should first lop a non-pathological dataset in the Urdu language to perform the speech recognition and then continue the effort for pathological datasets because voice disorders are highly effective psychological problems [2].

1. Urdu Language

The Urdu language is the official language of Pakistan. Urdu was hugely affected by Persian and Turkish and is written in a modified version of the Arabic script. The Persians adopted the Arabic letter in the 8th century, altering a few characters to represent Persian consonants found in Arabic. Some characters, such as the first two letters of the Urdu Alif and Alif MADD, may go directly beneath letters to alter the sound they make [9]. In table 1, five Urdu letters that have been chosen to form the Urdu speech dataset to conduct this study are shown with their written pronunciation.

Table 1. Pronunciation of selected Urdu letter in this study

Sr.	Letter	Pronunciation
1.	ا	Alif ('ā)
2.	ب	Bā'y
3.	غ	Ghayn [gh(ġ)]
4.	ك	Kāf
5.	ي	Yā' [y(ī.ay)]

2. Related Work

Al-Nasheri in Focus on an accurate and efficient method for detecting and classifying vocal disorders based on extracted features by studying the use entropy in various frequency bands of autocorrelation. The autocorrelation was used an objective of to collect the maximal peak and lag values from each spoken signal frame for disease identification and categorization. After normalizing his values as features, in addition, we calculated the entropy of the speech signal at each frame. To evaluate the contributions of each band to the sensing and classification process, these characteristics were examined across a range of frequency ranges. Several samples were extracted from three databases in for the continuation of a vowel, both for ordinary and pathological voices. The classifier was a support vector machine. If the averages of healthy and diseased samples vary considerably, then the U-tests were conducted. The best detection and classification accuracies achieved differ depending on the band, method, and database used. The most

significant bands were for both detection and classification, a frequency range of 1000 Hz to 8000 Hz is recommended. [10]. Further Al-Nasheri investigate the parameters of the Multidimensional Voice Program (MDVP) to detect, classify, and automatically classify the voice pathology in multiple data banks. The experimental results show a clear difference in the performance of these database MDVP parameters. The parameters highly ranked differentiated between databases. Three MDVP parameters adjusted in accordance with the Fisher discrimination rate yielded the highest accuracy and the most accurate parameters were obtained [11]. Lastly Al-Nasheri et, al. work focuses on developing a precise and robust extraction of features for determining, classifying, and investigating voice pathologies using correlation functions in different frequency bands. In the MEEI, they describe a new algorithm for exploring voice pathologies from the sounds of sustained vows, particularly in the light of a possible gap: the classification of the co-existing problems, which are the same as the principal phonic symptom, which implies similar interclass characteristics. Signal energy, null crossing rates, and signal entropy (SE) are used in the proposed technique to classify speech signals using the DPM, which provides an overview of the combined time and frequency data map [12].

Four diseases from the SVD dataset, including laryngitis, cyst, non-fluency syndrome, and dysphonia, were selected for analysis by Sidra et al. [13]. They extracted features from the audio signals and compared the results of four machine learning algorithms, including SVM, Nave Byes, decision tree, and ensemble classifier. They employed a comparison technique and a novel combination of features to identify laryngitis, cysts, non-fluency syndrome, and dysphonia in the SVD dataset using a purposeful sampling technique and the new features. To diagnose voice disorders with greater accuracy, a combination of particular 13 MFCC (Mel-frequency cepstral coefficients) characteristics, pitch, ZCR (zero-crossing rate), spectral flux, spectral entropy, spectral centroid, and short-term energy is used. An audio sample of 10ms has been shown to provide the best outcome when a mixture of characteristics is extracted. In the inter-classifier comparison, four machine learning classifiers, SVM (93.18%), Naive Bayes (99.45%), decision tree (100%), and ensemble classifier (51%), were used. Naive Bayes and the decision tree have the highest detection rates among these precisions. The suggested methodology's chosen collection of characteristics yields the best results using naive Bayes and decision trees. Furthermore, the SVM has been shown to be the most often utilized method for determining voice conditions. For the most part, clinical identification of voice abnormalities using machine learning algorithms has been the focus of most research, according to Sidra et al. As a result, we were able to improve the convolutional neural network's accuracy by 87.11 percent compared to the previously reported accuracy by the use of the suggested technique. The present neural network's accuracy is comparable to CNN's, and the implications were almost identical. Working with the SVD dataset's neural network for the identification of voice disorders will provide improved results in the future [14].

3. Dataset

The dataset collected to conduct this study contains non-pathological audio clips, i.e., a healthy voice at the data based on audio files that could serve the purpose of automatic speech recognition. There are a total of 37 letters in the Urdu language, out of which we have chosen five letters that can be seen in table 1. Depicting that there are many five classes in the proposed dataset. Each class contains 41 samples (each participant for every letter), so there are 205 recordings in the dataset. This dataset is still in the initial stages, adding more notes in the corpus. Dataset was prepared using the microphone of the iPhone 7 because iPhone has the best microphone installed that could easily filter the jittering in the recordings. Table 2 further represents the specifications of the proposed dataset.

Table 2. Specifications of the dataset that was developed in the proposed methodology

Specification Table	
Subject	Urdu language non-pathological voice dataset
Specific subject area	Computerized speech recognition with the help of a machine learning classifier
Type of data	Audio files
How data were acquired	Using a microphone of an iPhone
Data format	Raw, Analyzed
No. of classes	5 classes/letters
No. of samples	41 samples of each letter = $41 \times 5 = 205$
Parameters of data	Data was collected in a noise free environment
No. of participants	41
Demographics of participants	Forty-one participants include both men and women in between the range of 18 till 36 years.

4. Methodology:

In figure 1. SVM (Support Vector Machine) is used as a classifier to test the dataset collected in the first step of this paper. The training and testing dataset is divided into the ratio of 80% and 20%. MATLAB classification app is used to train the model. SVM is a decent tool for designing a speech recognition system. It tries to set up a threshold among classes, allowing labels to be anticipated from more than one vectors. This option, known as the hyper-plane, is chosen so that it is as far away as possible from the closest data points in each class. Support vectors are defined as the points that are closest to each other [15]. SVM classifier creates the most innovative hyperplane in the transformed entrance space, differentiates the exceptional groups, and maximizes the distance to nearby

cleanly separated instances. The variables of the hyperplane approach are a quadratic optimization problem [16]. To label a dataset, do the following:

$$(x_1, y_1) \dots (x_n, y_n), x_i \in R^d \quad (1)$$

Where x_i is a vector representation of a characteristic and y_i is a class mark (negative or positive) of a practice formula i . The optimal hyperplane is then described as follows:

$$wx^T + b = 0 \quad (2)$$

Where w denotes the weight matrix, x denotes the input vector, and b denotes the bias W and b must satisfy the following inequalities for all components of the training collection.

$$wx^T + b \geq +1 \quad \text{if } y_i = +1 \quad (3)$$

$$wx^T + b \leq -1 \quad \text{if } y_i = -1 \quad (4)$$

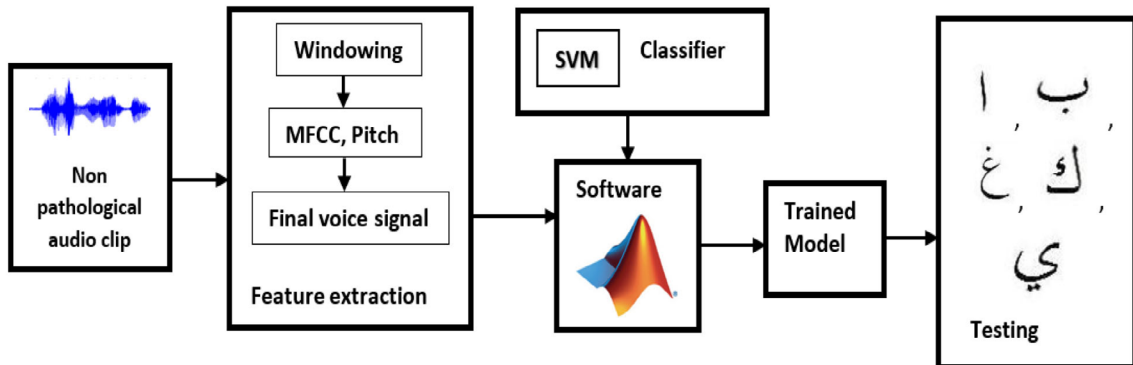


Figure 1. The proposed methodology that used two core features, namely MFCCs and pitch, are extracted from each audio clip and fed to an SVM classifier to predict 5 Urdu letters.

4.1. MFCC

The Mel frequency cepstral coefficient is influenced by the human auditory cortex. As per the perception research, the human auditory system does not work on a linearly inboud sound with an initial frequency 'f' recorded in hertz (Hz) and a pitch defined on the Mel scale [17]. MFCC is defined as coefficients deduced from audio signals. The voice input is the audio signal input that goes through the framing process. Before framing, the audio signal goes through a pre-emphasis process, which helps achieve accuracy and efficiency. This process compensates for the higher frequency suppressed in the human auditory

system throughout sound production.

$$C2(n) = c(n) - d * c(n-1) \tag{1}$$

Here, $c2(n)$ represents the output signal, and the recommended values for d are 0.9 and 1. The z transform of the filter is as follows:

$$H(Z) = 1 - D * Z^{-1} \tag{2}$$

Following the pre-emphasis process, the goal is to divide the entire audio signal into several frames so that each frame signal can be easily analyzed and interpreted. The audio signal is divided into 10 ms frames, whereas the standard framing size is 25 ms [18]. It demonstrates that the frame length for a 50 kHz audio signal is $50 \text{ k} * 0.01 = 500$ samples. The framing step allows the frames to overlap. There is a frame step of 10 ms for the first 500 samples. It starts at sample 0 and continues till the final audio signal has been heard, at which point the 500-sample structure ends. Since vowel recordings are included in the SVD dataset, the dataset specifies a recording duration of 10 ms, hence the 10 ms signal is used. When utilizing the Hamming window function, spectral artefacts may be minimized after framing. Convolution in the frequency domain results from the combination of short-term spectrum and window transfer function (hamming). Each frame must be multiplied with hamming window [19] to maintain continuity between the first and last marks in the frame. The hammering window function is described below.

$$K(n) = 0.54 + 0.64 \cos(2\pi nN - 1) \tag{3}$$

$K(n)$ denotes the window, and $Q(n)$ means the output, whereas $X(n)$ represents the input frame signal.

$$Q(n) = X(n) * w(n) \tag{4}$$

$$Q(w) = FFT[k(t) * X(t)] \tag{5}$$

$$Q(w) = k(w) * X(w) \tag{6}$$

The signal's original strength is now transformed to the Mel frequency using the Mel filter bank. Neither filter has a constant distance between them, nor is the number of filters in the higher frequency range less than those found in the lower frequency range. Filter banks are the only ones that can be used on signals in both the time domain and frequency domain. When processing Mel frequency cepstral coefficients in the frequency domain, it is important (MFCC). Figure 2 shows the filter applied in the lower and higher frequency regions to demonstrate the frequency change. Pitch and frequency may be linked by using the Mel scale. Low-frequency pitch variations may be distinguished from those occurring at higher frequencies by the human hearing system. [20] The Mel scale

is used to extract elements that are specific to human hearing. Using this mathematical technique, the frequency response may be converted into the Mel Scale:

$$M(f) = 1125 * \ln(1 + f/700) \quad (7)$$

Where f is the frequency of the audio signal.

5. Results:

In graph 1, the classifier's accuracy is calculated from the below formula. Whereas after finding the individual accuracies, the average accuracy is 85.866%.

$$AUC = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}$$

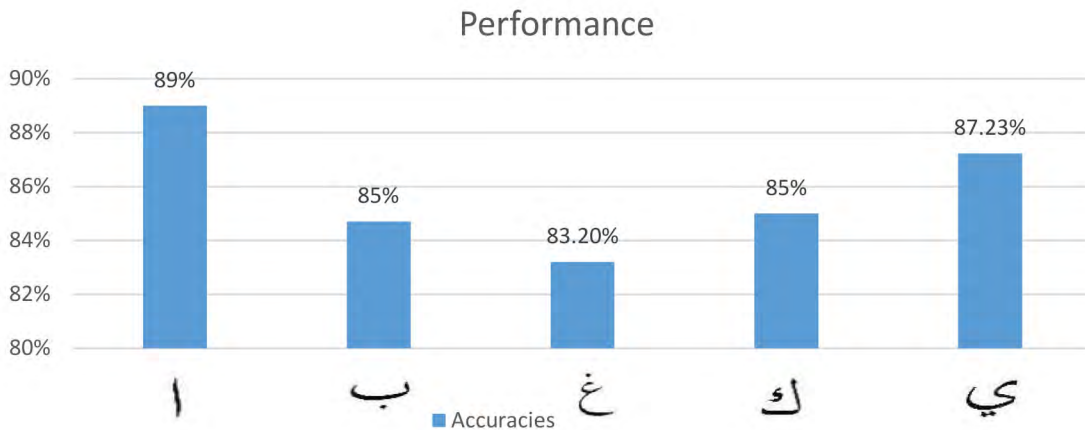


Figure 2. Accuracies of the voice signals of the Urdu letters

6. Conclusion:

Several studies have been conducted to detect voice pathologies. Because all cords are paralyzed, as a result of vocal cord paralysis, individuals often have a limited ability to speak or breathe. Invasive for patients, the screening test used to categorize these diseases of speech is intrusive in nature, so machine learning exploration and development have increased in recent years. The lack of analysis and automatic recognition of speech disorders of Urdu language encouraged authors to collect a non-pathological dataset. The average accuracy of the collected non-pathological dataset when trained and tested through the SVM classifier is 85.886%. In the future, we have planned to create a pathological dataset and perform automatic speech recognition by following the same method.

References

- [1] Graham Williamson. Human Communication: A Linguistic Introduction (2nd Edition) 2006.
- [2] ASHA Clinical Topics. Voice disorders. Website, 2019. <https://www.asha.org/PracticePortal/Clinical-Topics/Voice-Disorders>
- [3] Michael J. Clark James Hillenbrand, Laura A. Getty, and Kimberlee Wheeler. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(1):3099–3111, 1995.
- [4] T. Parsons. *Voice and Speech Processing*. McGraw-Hill College Div., Inc, 1986.
- [5] G. C. M. Fant. *Acoustic Theory of Speech Production*. Mouton, Gravenhage, 1960.
- [6] J. R. Deller, J. G. Proakis, J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.
- [7] D. O’Shaughnessy. *Speech Communication: Human and Machine*. Addison Wesley Publishing Co., 1987.
- [8] L. R. Rabiner, R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Inc., Englewood Cliffs, 1978.
- [9] "ATLAS - Urdu: Urdu Language", Ucl.ac.uk, 2021. [Online]. Available: <https://www.ucl.ac.uk/atlas/urdu/language.html>. [Accessed: 17- Sep- 2021].
- [10] Z. Ali et al., "Intra- and inter-database study for Arabic, English, and German databases: Do conventional speech features detect voice pathology?," *J. Voice*, vol. 31, no. 3, pp. 386.e1-386.e8, 2017.
- [11] A. Al-Nasheri et al., "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2018.
- [12] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," *J. Voice*, vol. 31, no. 1, pp. 3–15, 2017.
- [13] S. A. Syed, M. Rashid, S. Hussain, A. Imtiaz, H. Abid, and H. Zahid, "Inter classifier comparison to detect voice pathologies," *Math. Biosci. Eng.*, vol. 18, no. 3, pp. 2258–2273, 2021.
- [14] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative analysis of CNN and RNN for voice pathology detection," *Biomed Res. Int.*, vol. 2021, p. 6635964, 2021.
- [15] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (SVM) learning in cancer genomics, *Cancer Genomics-Proteomics*, 15 (2018), 41–51.
- [16] A. Shmilovici, Support vector machines, in *Data Mining and Knowledge Discovery*

- Handbook, Springer, Boston, MA, (2009), 231–247.
- [17] S. Memon, M. Lech, L. He, Using information theoretic vector quantization for inverted MFCC based speaker verification, in 2009 2nd International Conference on Computer, Control and Communication, IEEE, (2009), 1–5.
 - [18] M. Sahidullah, G. Saha, On the use of distributed data in speaker identification, in 2009 Annual IEEE India Conference, IEEE, (2009), 1–4.
 - [19] Ö. Eskidere, A. Gürhanlı, Voice disorder classification based on multitaper mel frequency cepstral coefficients features, *Comput. Math. Methods Med.*, 2015 (2015), 956249.
 - [20] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd edition, Wiley-Interscience, USA, 2000.


Call for Papers/Authors Guideline

KIET Journal of Computing & Information Sciences (KJCIS) is biannual publication of College of Computing & Information Sciences, Karachi Institute of Economics and Technologies. It is published in January and July every year. We are lucky to have onboard prominent and scholarly academicians as part of Advisory Committee and reviewers.

KJCIS is a multi-disciplinary journal covering viewpoints/ researches / opinions relevant to the non-exhaustive list of the topics including data mining, big data, machine learning, artificial intelligence, mobile applications, computer networks, cryptography and information security, mobile and wireless communication, adhoc and body area networks, software engineering, speech and pattern recognition, evolutionary computation, semantic web and its application, data base technologies and its applications, internet of things (IoT), computer vision, distributed computing, grid and cloud computing.

The authors may submit manuscripts abiding to following rules:-

- Certify that the paper is original and is not under consideration for publication in any other journal. Please mention so in case it has been submitted elsewhere.
- Adhere to normal rules of business or research writing. Font style be 12 points and the length of the paper can vary between 3000 to 5000 words.
- Illustrations/tables or figures should be numbered consecutively in Arabic numerals and should be inserted appropriately within the text.
- The title page of the manuscript should contain the Title, the Name(s), email address and institutional affiliation, an abstract of not more than 200 words should be included. A footnote on the same sheet should give a short profile of the author(s).
- Full reference and /or websites link, should be given in accordance with the APA citation style. These will be listed as separate section at the end of the paper in bibliographic style. References should not exceed 50.
- All manuscripts would be subjected to tests of plagiarism before being peer reviewed.
- All manuscripts go through double blind peer review process.
- Electronic submission would only be accepted at kjcis@pafkiet.edu.pk
- All successful authors will be remunerated adequately.
- The Journal does not have any article processing and publication charges.



Submission is voluntary and all contributors will find a respectable acknowledgement on their opinion and effort from our team of editors. Submission of a paper will be held to imply that it contains original unpublished work. In case the paper has been forwarded for publication elsewhere, kindly apprise in time if the paper has been accepted elsewhere. Manuscripts may be submitted before September and May to get published in Jan & July issues respectively. We encourage you to submit your manuscripts at kjcis@pafkiet.du.pk

Editorial Board KJCIS
College of Computing & Information Sciences
KIET Institute of Economics and Technology

Karachi Institute of Economics and Technology

Korangi Creek, Karachi-75190, Pakistan

Tel: (9221) 3509114-7, 34532182, 34543280 Fax: (92221) 35009118

Email: kjcis@kiet.du.pk

<http://kjcis.kiet.edu.pk>

KARACHI INSTITUTE OF ECONOMICS AND TECHNOLOGY

PUBLISHED BY:

College of Computing and Information

Sciences, KIET

cocis.kiet.edu.pk

kjcis.kiet.edu.pk

kjcis@kiet.edu.pk

Main Campus

PAF Airmen Academy,

Korangi Creek

Tel: 35091114 - 7

Cell: 0336-2508284 - 85

City Campus

Shahra-e-Faisal Site

28-D, Block 6, P.E.C.H.S.

Tel: 34546872, 34532182

Cell: 0336-2508286 - 87

City Campus

North Nazimabad Site

F-103 Block F, N. Nazimabad

Tel: 36628381, 36679314

Cell: 0336-2444191 - 92