



Center for Open Access in Science

Open Journal for
Information Technology

2023 • Volume 6 • Number 2

<https://doi.org/10.32591/coas.ojit.0602>

ISSN (Online) 2620-0627

OPEN JOURNAL FOR INFORMATION TECHNOLOGY (OJIT)

ISSN (Online) 2620-0627

www.centerprode.com/ojit.html * ojit@centerprode.com

Publisher:

Center for Open Access in Science (COAS), Belgrade, SERBIA

www.centerprode.com * office@centerprode.com

Editorial Board:

Phynos Mylonas (PhD)

Ionian University, Department of Informatics, Corfu, GREECE

Petra Grd (PhD)

University of Zagreb, Faculty of Organization and Informatics, CROATIA

Silvia Nikolaeva Gaftandzieva (PhD)

University of Plovdiv "Paisii Hilendarski", Faculty of Mathematics and Informatics, BULGARIA

Biserka Yovcheva (PhD)

Konstantin Preslavski University of Shumen, Faculty of Mathematics and Computer Science, BULGARIA

Alen Kišić (MSc)

University of Zagreb, Faculty of Organization and Informatics, CROATIA

Andrey Fomin (PhD)

Saratov State University, Faculty of Nano and Biomedical Technologies, RUSSIAN FEDERATION

Executive Editor:

Goran Pešić

Center for Open Access in Science, Belgrade, SERBIA

CONTENTS

- 67 Design and Implementation of a Tailoring Management System (Virlor)
Adebola Victor Omopariola, Raphael Ozighor Enihe, Chukwudi Nnanna Ogbonna, Felix Uloko, Manasseh Chukwudubem Ezeocha & Monday J. Abdullahi
- 97 Automated Assessment System Using Machine Learning Libraries
Victor Adebola Omopariola, Chukwudi Nnanna Ogbonna, Felix Uloko & Monday Abdullahi
- 123 Detecting Phishing Emails Using Random Forest and AdaBoost Classifier Model
Fredrick Nthurima, Abraham Mutua & Waithaka Stephen Titus
- 137 Managing the Implementation of Information Technology in Schools
Alaa Sarsour & Raed Sarsour
- 143 A Hybrid Model for Detecting Insurance Fraud Using K-Means and Support Vector Machine Algorithms
Brian Ndirangu Muthura & Abraham Matheka
- 157 A Classifier Model to Detect Phishing Emails Using Ensemble Technique
Fredrick Nthurima & Abraham Matheka





Design and Implementation of a Tailoring Management System (Virlor)

Adebola Victor Omopariola, Raphael Ozighor Enihe, Chukwudi Nnanna Ogbonna,
Felix Uloko & Manasseh Chukwudubem Ezeocha

Veritas University, Abuja, NIGERIA
Department of Computer and Information Technology

Monday J. Abdullahi

Air Force Institute of Technology, Kaduna, NIGERIA
Department of Computer Science

Received: 4 January 2023 ▪ Revised: 19 April 2023 ▪ Accepted: 7 May 2023

Abstract

The tailoring method has a popular view of being a Manuel method, in which clients must attain out to the tailor physically, select the materials for their clothing, provide their measurements, and, in most circumstances, return to the tailor shop to pick it up, consuming more time and more resources. as time progresses the provision of services has modernized, and even preference of service provision of customers has also modernized as well, customers of recent would prefer to employ mobile services that can easily reach them rather than Manuel conventional service deliveries. in this recent society, the majority would prefer a service that is automated to the point they put little or no effort into acquiring these serviced thus the goal of this project, the project is aims to automate the tailoring management services which is manually maintained. After the automation this will mean better services, data security, quick search, and also paperless environment. The project's major goal is to automate these services in such a way that tailors will have more work opportunities. Every user of the system will have to log into the system using username and password so that security and authentication will be ensured. After logging in, a consumer can place an order, monitor the status of their outfit, and even provide feedback. This system would aid both tailors and customers as it improves effectiveness and efficiency, this system also help in realizing the vision 2030 where the Nigerian government wants its people to be digitally informed and automate all government bodies and ministries, thereby embracing Electronic Governing.

Keywords: tailoring management system, design, implementation.

1. Introduction

The Virlor tailoring service application is a system which automates the tailoring process, creating job opportunities for tailors, and improve tailoring service delivery. it provides services to tailors such as measurement submission to tailor, check whether their garments are completed, as well as an assessment bar provided to clothiers to reflect the number of successfully completed jobs/tasks. The goal of the research is to create and build a virtual tailoring management system that is more effective and efficient than the current manual technique.

1.1 *Background of study*

Tailoring is the art of designing, cutting, fitting, and finishing clothes. The word tailor comes from the French “*tailer*”, to cut, and appears in the English language during the fourteenth century. In Latin, the word for tailor was *sartor*, meaning patcher or mender, hence the English “sartorial,” or relating to the tailor, tailoring, or tailored clothing. The term bespoke, or custom, tailoring describes garments made to measure for a specific client. Bespoke tailoring signals that these items are already “spoken for” rather than made on speculation.

As a craft, tailoring dates back to the early Middle Ages, when tailors’ guilds were established in major European towns. Tailoring had its beginnings in the trade of linen armorers, who skillfully fitted men with padded linen undergarments to protect their bodies against the chafing of chain mail and later plate armor. Men’s clothing at the time consisted of a loosely fitted tunic and hose. In 1100 Henry I confirmed the royal rights and privileges to the Tailors of Oxford. In London, the Guild of Tailors and Linen Armorers were granted arms in 1299. They became a company in 1466 and were incorporated into the company of Merchant Tailors in 1503. In France, the tailors of Paris (*Tailleurs de Robes*) received a charter in 1293, but there were separate guilds for Linen Armorers and Hose-Makers. In 1588, various guilds for French tailors were united as the powerful *Maitres Tailleurs d’Habits*. Tailoring has traditionally been and remains a hierarchical and male-dominated trade, though some women tailoresses have learned the trade. (Mathew, 2022).

In recent times, tailoring has been viewed as a profession for the unlearned, especially in the Nigeria society, it is seen as a profession for school dropouts and those who aren’t privileged to complete their schooling. This is simply an effect of how conventional the means of carrying out tailoring services have been. Clients must travel considerable distances to obtain their measurements, and afterwards, measurement is then recorded on a sheet of paper which can easily be misplaced or damaged. These methods pose threat to the security of customers and even the information recorded. On the tailor’s end, most new tailors open stores and barely get customers due to the fact that most customers are not aware of their business. The virtual tailoring system would solve all this problem and provide an environment where these activities are carried out more efficiently and effectively.

1.2 *Problem statement*

A common problem experienced by customer’s is measurements, customers have to walk to the tailor shop just to get their measurements recorded, not only that being a case, the measurements are recorded on a sheet of paper, customers also need to constantly go to the tailor’s shop to find out if their attires done. Tailors also find it difficult to acquire shops that would be used for their business. The manual system in use has too many setbacks that make the tailoring service difficult.

1.3 *Proposed solution*

The Virilor tailoring service system would eliminate each of the problems listed above, the system would allow customers submit their measurement online, the system would also keep measurements recorded, this record would be edited by the tailor consecutively due to the fact that the human body changes. The system would also display the garment status indicating when its suitable to be picked up. The system would also display the cost of the garments so customers can have prior knowledge about it.

1.4 *Project aims and objectives*

This project aims to automate the manual tailoring system service that has been seen to be tedious, discomforting and tiring. It has the following objectives:

- This system also would enable the sending of measurements online.
- This system also aims to compute the price of a garment prior to the making phase, including price of fabric and price of style preparation.
- This system possesses an ordered organization format which makes tasks easier to pick by tailors in an orderly manner.

This system aims to modernize the entire tailoring process.

1.5 *Scope of system*

The Virlor system will provide an interface for users to register and share their measurements and other processes are as follows:

- The system stores and maintains clients/customers recorded measurements which can be edited from time to time.
- The Virlor system also provides prior price statements totaling the amount spent on fabric, sewing and delivery.
- The choice is done by the client, choosing to either transfer payment or pay on delivery.

For the purpose of this system, we will have only two of users to make it effective and secures. The users will be the clients and the admin.

administrator (user/tailor)

- have access to register to the Virlor application.
- have access to checking cloth status.
- have access to tailor navigation interface.
- can login to change application interface.
- can magnate the information on the site.
- can remove tailor based on feedback gotten from clients.

1.6 *Scope of the study*

The scope of this research, design and development dwells on Abuja, Nigeria and a few prominent locations surrounding it. This is because the Manuel system is practicalized worldwide, in every country the Manuel tailoring system is adopted, and due to the lack of research time restriction, lack of needed research materials and resources, the scope of the study is limited.

1.7 *Research question*

The purpose of my project topic was inspired by the following research question: *Can design and implementation of a tailoring management application improve the tailoring process?* This question would be answered in our Chapter 3.

1.8 *Significance to the study*

This study has its relevance in solving real life problems affecting tailors/clients today ranging from measurements, fabric selection, and general organization for the tailoring process. The application would also be flexible to provide different tailoring options where users can have specific choices to select from. Lastly the application would be focused on one state i.e., federal Capital Territory rather than addressing a greater audience. It is important to put focus interest into the development and usage of the modernized tailoring management system in solving tailoring issues.

The major limitation of this study is inadequate time due other rigorous academic work that have to embark on as a final year student at the course of study.

1.9 *Operational definition of terms*

(1) Tailoring: Tailoring is the art of designing, cutting, fitting, and finishing clothes.

(2) Tailor: A patcher, designer, or mender.

(3) System: A set of things working together as a whole to achieve the same sole objective.

2. Literature review

2.1 *Introduction*

Literature review is a text written by someone to consider the critical points of current substantive results as well as theoretical and methodological contributions to a particular issue are all examples of knowledge. The main goals are to place the current study in the context of the entire literature and to provide context for the reader (Cooper, 2020).

A literature review is a piece of academic writing that demonstrates knowledge and comprehension of academic literature on a given topic in context (Rudestam et al., 2019).

2.1.1 *Who is a Tailor?*

A tailor is a person who makes, repairs, or alters clothes professionally, especially suits (Yourdictionary, 2019). A tailor creates bespoke clothing, such as jackets, in a variety of styles. Skirt or trousers that go with them, for men or women. An alteration specialist which adjusts the fit of garments (Lancaster, 2019).

2.2 *Origins of the term “bespoke tailoring”*

According to Poole (1846), the term bespoke originated in the days when a customer would select a chunk of cloth in a tailor’s store and have it marked as “bespoken for” by the tailor. It has come to represent a classic kind of tailoring in which each client’s pattern is created individually and the best traditional tailoring skill is employed to get the final garment’s shape. The following are the two main reasons for bespoke designed clothing:

1. Difficulty attaining a good fit from ready-to-wear garments.
2. Access to a wider range of styles and cloth designs.

According to Hardy (2020), a skilled tailor should be able to overcome all potential flaws and steer his customer toward a style that is more suited to his or her physique, as well as create a masterpiece that fits. A good tailor, he claims, can construct simple clothing from plain material, but with time and effort, they may learn to create garments of amazing beauty that provide considerable protection to their wearer (Hardy, 2019).

2.3 Developments in tailoring industry

Nigeria's fashion sector has increased in size and complexity over the last decade, gaining international notice. The “textile, garment, and footwear” sector has grown at an annual rate of 17 percent since 2010, according to GDP data from the national bureau of statistics (NBS). The rise has been spurred in part by increased demand, but it has also been fueled in part by unprecedented initiatives that have pushed Nigeria into the global fashion awareness (Ogunfuyi, 2019).

He brought a rebellious streak to the history of suit creation, according to Richard (2019), and he has become a pillar of the modern menswear establishment. The rock n' roll elite have found his bright color and inventive twists alluring.

Within Savile Row in London, modernization of the style and attitude of traditional tailors to new designs has resulted in greater profitability, time wastage, and a reduction in the number of tailors that rely on traditional technology (Ozward et al., 2019).

According to Shaw (2001), the only man who behaves reasonably in his tailor shop is the one who takes fresh measurements every time he sees me, while the rest keep their old measurements and expect me to suit them (Shaw et al., 2020).

2.3.1 Existing tailoring systems

David Mutembei a graduate of the MT Kenya University built a tailoring management system in 2013, these systems consist of three major actors in the system which were: the users, the tailors, and the system admin.

But this differ from my system because my system has one major actor which is the tailor, he is the admin of the system, and manages the general process of the system.

The system will allow customers to register online and successfully submit their measurements.

The system has inbuilt validation system to validate the entered data. The customer can login to the system to check on the status of the clothes for collection. The system will show the already completed garments for clients to collect. The system also provides information about the cost of each garment the customer intends to get knit. This data will be stored in the database for further reference or audit.

Cletus (2020) developed a three-actor tailoring system which automates the tailoring process, but significant to his work, his system didn't request customer feedback after a successful tailoring process, his project focused on an Online tailoring management system aimed to assist in management of tailoring activities within the industry. It will provide online services to customers such as: measurement submission to their tailors, check whether their garments are finished and also help in proper keeping of records. This will ensure availability of right information, information safety, easy storage, access and retrieval. The study aims at building a computerized tailoring management system that would be more effective and efficient than the existing manual system.

This proposed online tailoring management system aimed to eliminate all these manual interventions and increase the speed of the whole process. His system also allows customers to register online and successfully submit their measurements.

His system contained an inbuilt validation system to validate the entered data. The customer can login to the system to check on the status of the clothes for collection. His system will also show the already completed garments for clients to collect. The system also provides information about the cost of each garment the customer intends to get knit. This data will be stored in the database for further reference or audit.

Igbes Online Tailoring management system also broke the geographical barriers and bring the whole process into a quick and easy way to access tailors. His system also automates the traditional tailoring system into a modern computerized system. This will enhance data retrieval, storage, and security. His system is also cost effective since it will cut down on travelling cost to get your measurements taken and also going to check if you clothe has been made and is ready for collection.

Another existing system was found to be completely manual, i.e., customers' information is captured in books, there also required to walk to the tailor shop to get their measurements taken.

Customers also go to the tailor shops to check on the progress of their garments.

2.3.2 Becoming a twenty-first century tailor shop

Smaller bespoke producers have been able to re-envision profitable business processes to reach global audiences because to the growth of online retail and advances in web technologies. With the development of collaborative digital market places like StanfordRow.com, bespoke industries are seeing a strong return. Despite the fact that a bespoke suit was “fully handmade and the pattern cut from scratch, with an intermediary baste stage which involved a first fitting so that adjustments could be made to a half-made suit, both fully bespoke and made-to-measure suits were “made to order” in that they were made to the customer’s precise measurements and specifications, unlike off-the peg suits” (Michael et al., 2019).

2.3.3 Moving online

While many older businesses are hesitant to adopt technology-driven business strategies, younger entrepreneurs are gaining market share by utilizing technology on numerous fronts.

Startups can use distance tailoring to expand their reach beyond the geographic boundaries of their local market. Customers measure themselves (with assistance) and submit orders online. Although many tailors utilize this method to take advantage of cheap labor overseas, a remote tailoring framework can also be used (gaebler.com, 2019).

Distance tailoring. Distance tailoring allows startups to expand their reach beyond the geographic limitations of the local marketplace. Customers perform their own measurements (with guidance) and place orders online. Although many tailors use this approach to take advantage of cheap labor overseas, it's possible to leverage a distance tailoring framework. (gaebler.com, 2020)

Integrated backend solutions. Tailor shops are like any other SMB (small and medium business) in the sense that there are multiple behind-the-scenes business tasks that must be routinely performed. With today’s technology, accounting, billing, inventory, shipping, and other

software solutions can be integrated to create a highly functional and seamless backend system. (gaebler.com, 2022)

Social media marketing. Social media resources like Facebook and Twitter allow tailor shop startups to convert satisfied customers to brand advocates. By actively engaging your customers on these and other sites, you can encourage positive conversations around your products and your brand (gaebler.com, 2022).

2.3.4 Web 2.0 technology in tailor systems

Web 2.0 is a participatory platform, through which consumers can download content, as well as contribute and produce new content by uploading. There are more ideas linked with this technology such as tagging, blogs, wikis, and mashups which link both retailers and consumers. Web 2.0 fashion product viewing and service technologies have advanced significantly and have been in use since the fashion industry joined e-commerce platforms (Idrees, 2020).

Web 2.0 Fashion Product Viewing Technology

Web 2.0 fashion product viewing technology is acknowledged and available online. It is described as a method of visual merchandising in online atmosphere. Viewing products online offers consumers basic knowledge related to the product and facilitates the decision to make a purchase.

(a) 2D Image Viewing To views a product and its features in online fashion retailing, 2D images are commonly seen on a model view or as an outfit view. The 2D product images provided by retailers increase the consumer's intention to purchase a product. The 2D images are not lively and are less engaging than the extensive viewing technology of zoom and product video display. Moreover, styling inspiration is also provided by the model wearing an outfit allowing consumers to imagine themselves wearing the product. Product video is a more charismatic tool for gilded sensory visualization of the product's characteristics than 2D images. Consumer's intention to purchase is empowered by presenting more sensory features in online platforms. Thus, utilizing charismatic media tools can boost consumers' engagement and probability of purchasing product (Idrees, 2020).

(b) Front, Side and Back Viewing There is an increased opportunity to sell a product by displaying images taken from various angles, which provide more information for consumers to evaluate before buying a product. The consumer is aided with multiple images of a product which enhance the sensory empowering experiences, thus making up for the absence of palpability in fashion ecommerce. The mental perceptibility increases proportionally to the increase in number of images displayed. Consequently, purchase intention is enhanced (Idrees, 2020).

(c) Angled Viewing There is an option of angled viewing in web 2.0, which allows consumers to view the product from various perspectives and with in-depth detail and information. The interactivity and engagement of the consumer increase by viewing the product at different angles. Mental tangibility increases by imagining how the product might look once purchased. With the angled viewing tool, consumers can operate the image they are viewing online. Decision making attribute is enhanced by the provision of a higher level of product involvement (Idrees, 2020).

(d) Zoom (Close-UP View) Zoom provides an option for involvement within the online retail setting, facilitating online consumers with enhanced product presentation and task-related knowledge to purchase a product. Conversely, there are so many online product viewing tools that, determining which is the most (Idrees, 2020).

The augmented product on their body on a digital screen. The augmented technologies included two basic categories:

Augmented 3D Product View and Virtual Mirror

(1) Augmented 3D Product View: Augmented reality delivers a higher level of experiential value during online shopping. AR is a new interactive technology through which consumers can interact with an augmented product over the real image of a person. This technology creates a seamless interactive environment between a person and the viewed product. This field is still under research (Vignali, 2020).

(2) Virtual Mirror: Virtual mirrors deliver a greater level of manifestation than 360-degree spin and motionless images. It is suggested to incorporate virtual mirrors in online retailing, which will act as a medium to minimize gap between the online and offline environments (Vignali, 2020).

Virtual Technologies Virtual reality (VR) is associated with cyber technology. VR is a virtual manifestation experience, although not the experience of direct occurrence, but rather the feeling of being engrossed in the virtual atmosphere. Virtual reality is the human perceptibility and the conversation of manifestation. VR technology can incorporate virtual avatars in the virtual fitting rooms of e-commerce platforms. The virtual technologies included four main types in e-commerce platforms such as (Vignali, 2020): (1) Avatars to mix and match for dynamic product view; (2) Virtual fitting rooms; (3) Virtual catwalk; and (4) Virtual Body scan.

(1) Avatars: 3D avatars enhance entertainment qualities for users and encourage consumers to revisit a website. Consumers can mix and match products on avatar for dynamic product view. Formerly, information was limited to product-related features. The consumer's decision-making is highly dependent upon information provided by consumers at this search stage. This technology is a consumer-oriented technology. Shim and Lee in 2011 determined that, avatars permit a greater level of telepresence. Telepresence provides a feeling of physical atmosphere and is described properly and efficiently as a place where users take part with interactive experience. Research suggests that telepresence is enhanced by 3D models. Virtual models can also enhance perceived enjoyment and hedonic value as well as perceived ease of use and usefulness (Gill, 2020).

(2) Virtual Fitting Rooms (virtual try-on): In contrast with traditional retailing setups, virtual fitting rooms improve the experience of consumers with innovation and curiosity. Curiosity is encouraged with virtual fitting rooms which also enhances the probability of both online and offline store benefaction. During customers' transactions, engaging with virtual fitting rooms can also enhance and refine the retailer's targeting strategy of collecting users' data for providing a personalized service. Moreover, virtual fitting rooms can improve consumers' empirical attitude. Purchase intention and functional value improves with personalized virtual try-on (Gill, 2020).

(3) Virtual Catwalk: During the launch of new seasonal products virtual catwalks are frequently used by technology-oriented fashion brands. The virtual reality headsets are used to present a virtual catwalk. Oculus Rift, Samsung Gear and more recently Google Cardboard are big brands names included in the list. Moreover, virtual catwalk is presented by Topshop during fashion week, as well as Burberry and Rebecca Minkoff (Gill, 2020).

(4) Virtual Body Scan: Professional body scanning equipment is used for 3D body scanning. Positive responses have been concluded by consumers when using 3D body scanning technology. Latest mobile applications have been launched by various companies. Size stream at home is the latest mobile application for body scanning. This technology provides measurements of the human body by taking images of a person wearing a specific suit on its application. This scanning technology was introduced due to inaccuracy in manual measurement methods used by

consumers for customization (Gill, 2020). There is also the Intelligent Web 3.0 Fashion Technology.

3. Methodology

The term methodology means the techniques and procedures adopted by conducting a research study. It outlines how the data will be collected, and the tools for collecting data, system methodology, the proposed system input and output, users and system development tools.

3.1 *Fact finding techniques*

This shows how data will be collected from the users of the Virlor system. The techniques used for data collection are as follows.

3.1.1 *Observation*

This technique would be used to collect data from users by carefully examining the processes carried as users interact with the Virlor system, and interacting with users who use the Manuel tailoring method. This involves systematically watching and recording the behaviors and characteristics of operations and processes. It gives more detailed and context related information and can adapt to events as they occur, however, the method may be time consuming.

3.1.2 *Secondary data collection*

This method of data collection involves data being collected from already existing works such as books, websites, magazines, newspapers, works gathered and analyzed by other fellow researchers, the research would then be compared with primary data collected and final decisions and conclusion would be made.

3.2 *System development methodology (SDLC)*

System development methodology is a technique that is used to show how the proposed system will be developed. In this case, the methodology used will be a waterfall model.

3.3 *Waterfall model*

It is made up of the stages that the programmer will go through when creating the system. The name waterfall comes from the fact that it is a sequential model. Before moving on to the next level, the developer must complete the previous one. The feasibility study, analytical phase, design phase, coding phase, testing phase, implementation phase, and maintenance phase are all included. It's a straightforward model that's simple to use and comprehend. With waterfall development approaches, analysts and users move from one phase to the next in a logical order. As the project progresses from phase to phase, the deliverables from each phase become increasingly large and are delivered to the project sponsor for approval. The phase concludes when the sponsor approves it, and the following phase begins.

Diagram of Waterfall model

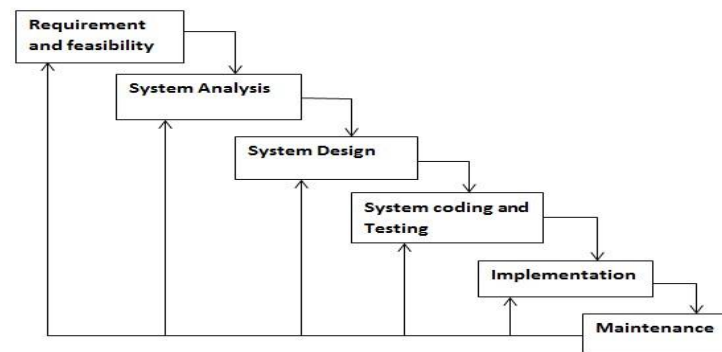


Figure 1.

3.4 Feasibility study

Through interviews, observations, and obtaining secondary information, we will conduct a study to gain a better knowledge of the customers' present system and the challenges they are experiencing with it. We'll utilize the information we've gathered to assess the technical, economic, and social feasibility of the system under consideration.

3.5 Requirements analysis

At this point, we'll gather information about the customer's requirements and outline the challenges that the system will be expected to solve. we'll also provide information on the customers' businesses, product features, and compatibility. we'll gather information about software, such as the programming language to be used, the database model to be used, and the gear required, such as laptops and printers.

3.6 Design

An overall design of the system architecture and physical design, which includes the user interface and database design, is completed at this stage. Before moving on to the next stage, we will identify any flaws at this point. The design specification is the stage's result, and it's used in the next step of implementation.

3.7 Coding/Implementation

At this point, we'll start coding according to the design specifications. This process produces one or more product components that have been debugged, tested, and integrated to meet the system architecture requirements using a pre-defined coding standard.

3.8 Testing

At this step, we'll make sure that both the individual parts and the entire system are thoroughly tested to guarantee that they're free of errors and meet client expectations. We'll include unit testing of individual code modules, system testing of the entire solution, and customer acceptance testing. Before moving on to the next stage, we'll make sure all bugs are fixed. At this time, we'll also be preparing, reviewing, and publishing Product documentation.

3.9 Maintenance

Once the product has been evaluated and verified as safe to use, it is then packaged. The system is ready to be installed at the customer's location. Depending on the demands of the consumer, we will distribute via the internet or by mail.

This stage takes place once the installation has been completed. It entails making changes to the system in order to increase performance. These updates are either user-initiated or as a result of previously unknown problems being uncovered. These changes are documented and the system is updated as a result of them.

Using the requirements definitions as a foundation, the system design is now constructed. Software design is the process of representing the functions of each software system in a manner which may readily be transformed to one or more computer programs. Use case diagrams, sequence diagrams, entity relationships frustrate (ERD), data dictionary and so on are used to this level to represent the system design.

3.9.1 Use case diagram

The use case diagram of the actor and different cases. A depiction of a systems behavior or functionality under various conditions as the system responds to requests from users. In the general use case, the summary of the fictional requirement is given. This shows the relationship of all users and the various cases involved.

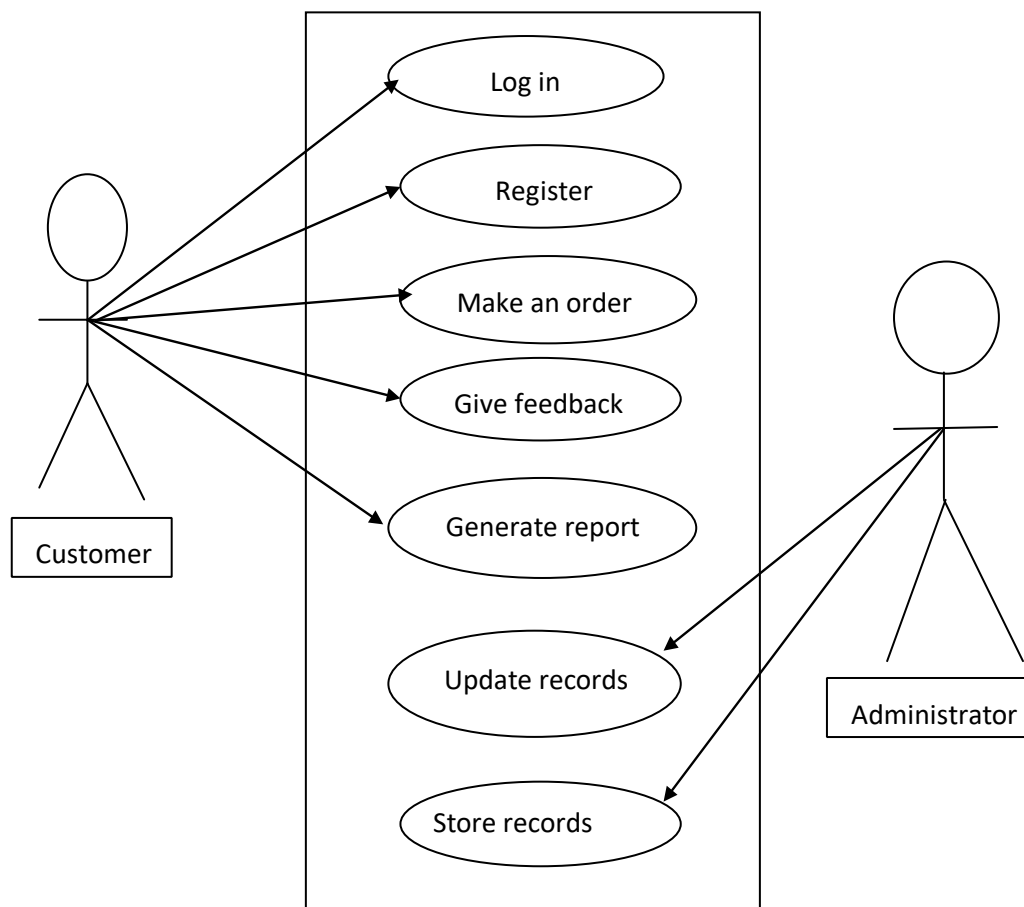


Figure 2. The use case diagram of the proposed system

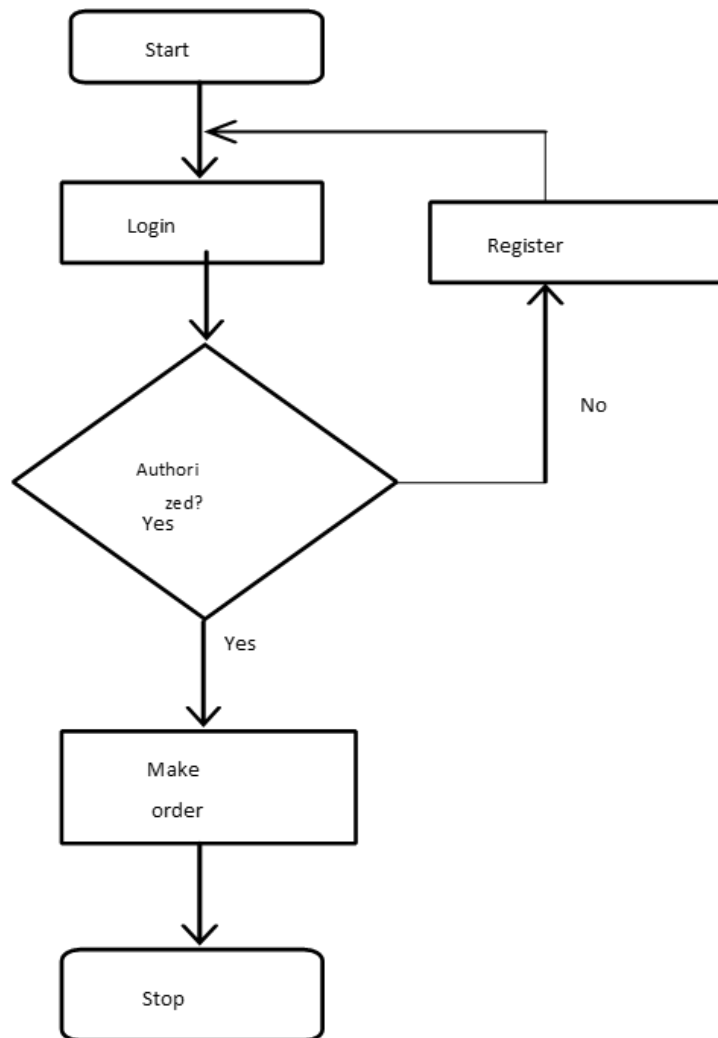
3.9.2 Input design

The input part of the system depicts just one section, which is the admin of the system, the admin of the system are the users of the website. These pages would be created and implemented by encoding them with PHP. the user is able to interact with the system by clicking the different functions of the site from placing delivery, to receiving delivery pickup information.

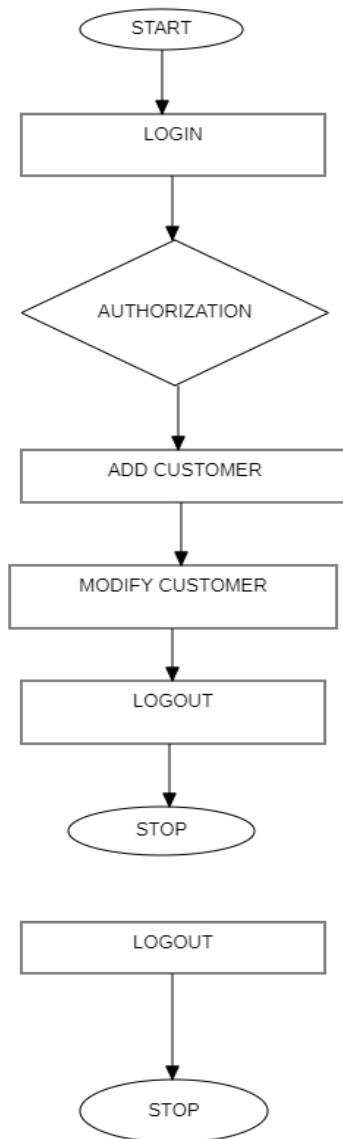
3.9.3 Proposed system

The proposed system is mainly classified into one important part, the administrative sector interface. The administrative sector is the back end of the system which tidies the functionality of the system. This part been handled by the system administrator. Its sole objective is to feed the system with up-to date information about the tourism centers in the state and every other important information. When the administrator logs in with his identification password, he is able to carry out his duties which ranges from the addition of newly discovered tailors, uploads new feedback scales, every other necessary information user's may need. It is the backbone of the system.

Login operation



Flowchart



4. System implementation, testing and integration

4.1 Introduction

It is the processes of putting the proposed system in operation. Some of the activities undertaken by the analyst are training personnel who will use the system. There is also provision of user manual and help page for efficient use of the system.

Next is to install computer equipment and internet to help them connect with their clients in the globe. This will facilitate the full functionality of this proposed system. Equipment should be acquired from recognized vendor. These include central processing unit (CPU), Ethernet cables, routers, output, and input devices, e.g., keyboard, mouse, monitor and all secondary storage devices. The hardware and software vendors have major responsibility for installing this

equipment. The analyst then determines the functional changes. E.g., may analyze the job function changes caused by the computerized system.

The software development life cycle (SDLC) ensures that before a product is deployed, it is tested and integrated to ensure proper functionality. System testing, integration and implementation are phases in the SDLC, which ensures that the aim of the system development life cycle is achieved before the release of any product.

Ogwuleka (2012) defined system implementation as the construction of a new system and the delivery of the system into

4.2 The choice of programming language

The project which is a based system, requires the use of a couple of selected programming languages which include PHP and Perl programming languages to aid the implementation of the Virlor system. These languages fall under the programming paradigm called scripting languages.

PHP is a general-purpose scripting language geared toward web development. It was originally created by Danish-Canadian programmer Rasmus Lerdorf in 1994. The PHP reference implementation is now produced by the PHP group. PHP originally stood for personal home page, but it now stands for recursive initialism PHP: hypertext preprocessor.

In accordance with Kalode (2021), PHP is an open-source server-side scripting language that many developers use for web development. It is also a general-purpose language that you can use to make lots of projects, including Graphical User Interfaces (GUIs).

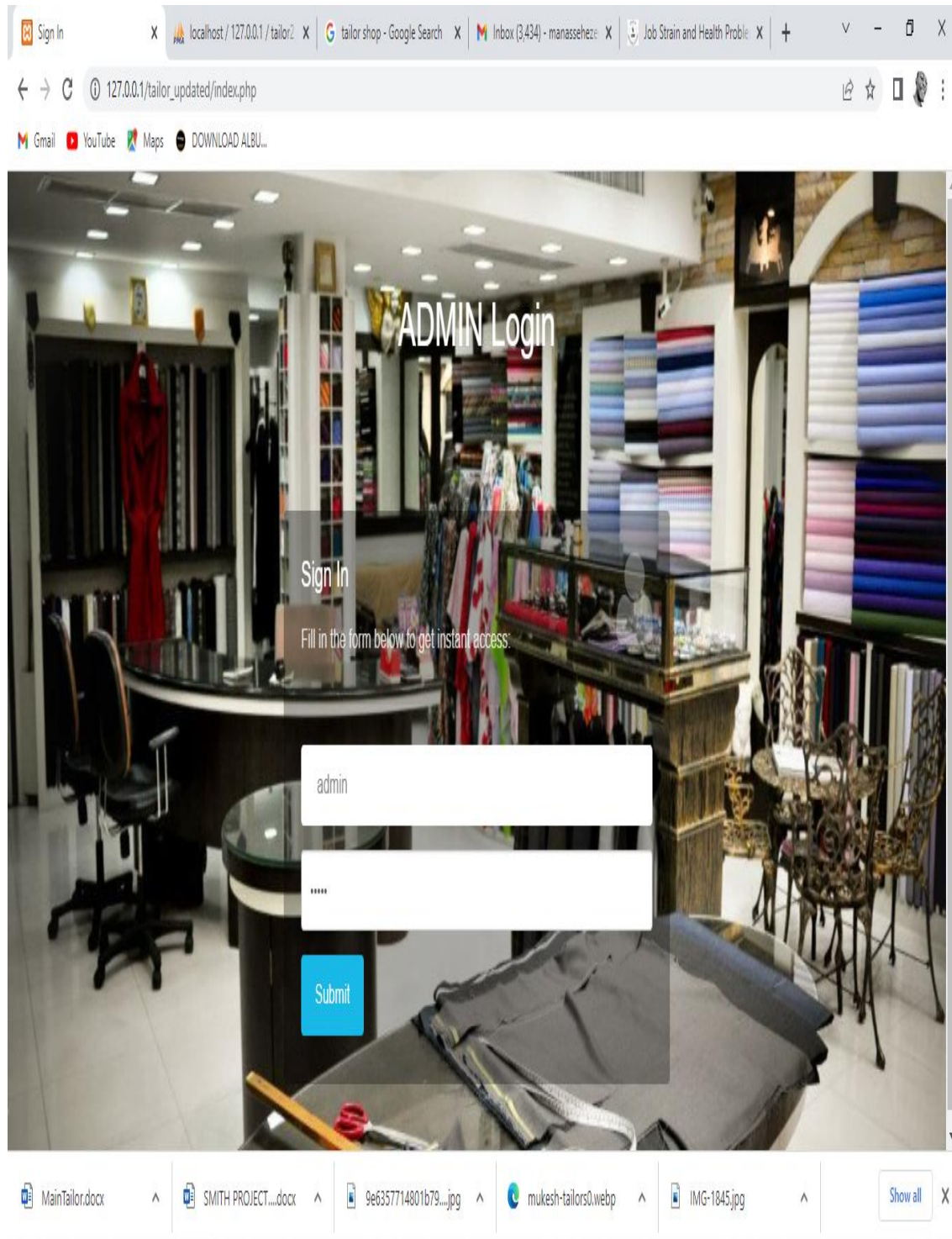
The abbreviation PHP initially stood for Personal Homepage. But now it is a recursive acronym for Hypertext Preprocessor. (It's recursive in the sense that the first word itself is an abbreviation, so the full meaning doesn't follow the abbreviation.)

The first version of PHP was launched 26 years ago. Now it's on version 8, released in November 2020, but version 7 remains the most widely used.

PHP runs on the Zend engine, which is the most popular implementation. There are some other implementations as well, like parrot, HPVM (Hip Hop Virtual Machine), and Hip Hop, created by Facebook.

PHP is mostly used for making web servers. It runs on the browser and is also capable of running in the command line. So, if you don't feel like showing your code output in the browser, you can show it in the terminal.

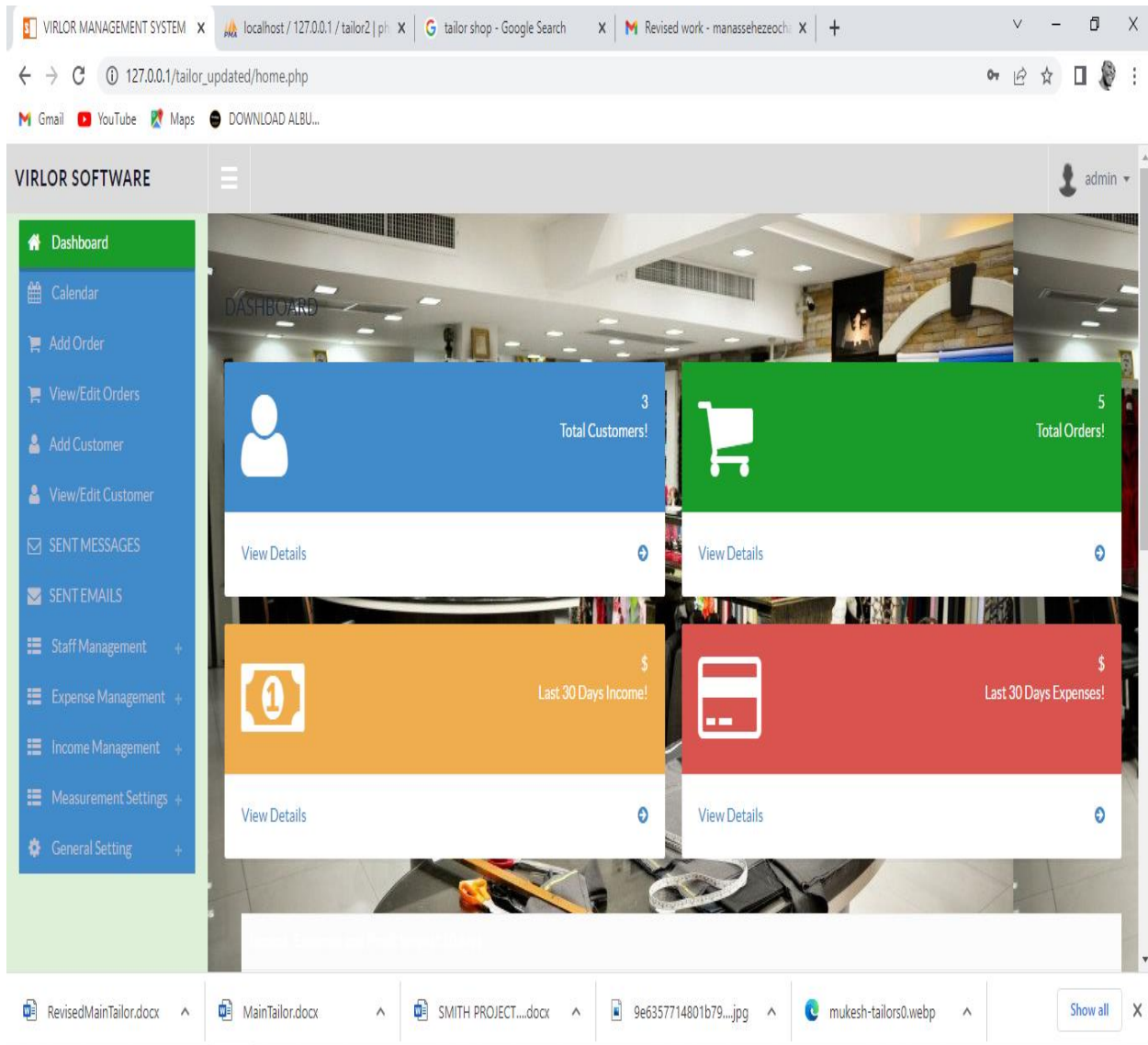
4.3 System main menu implementation



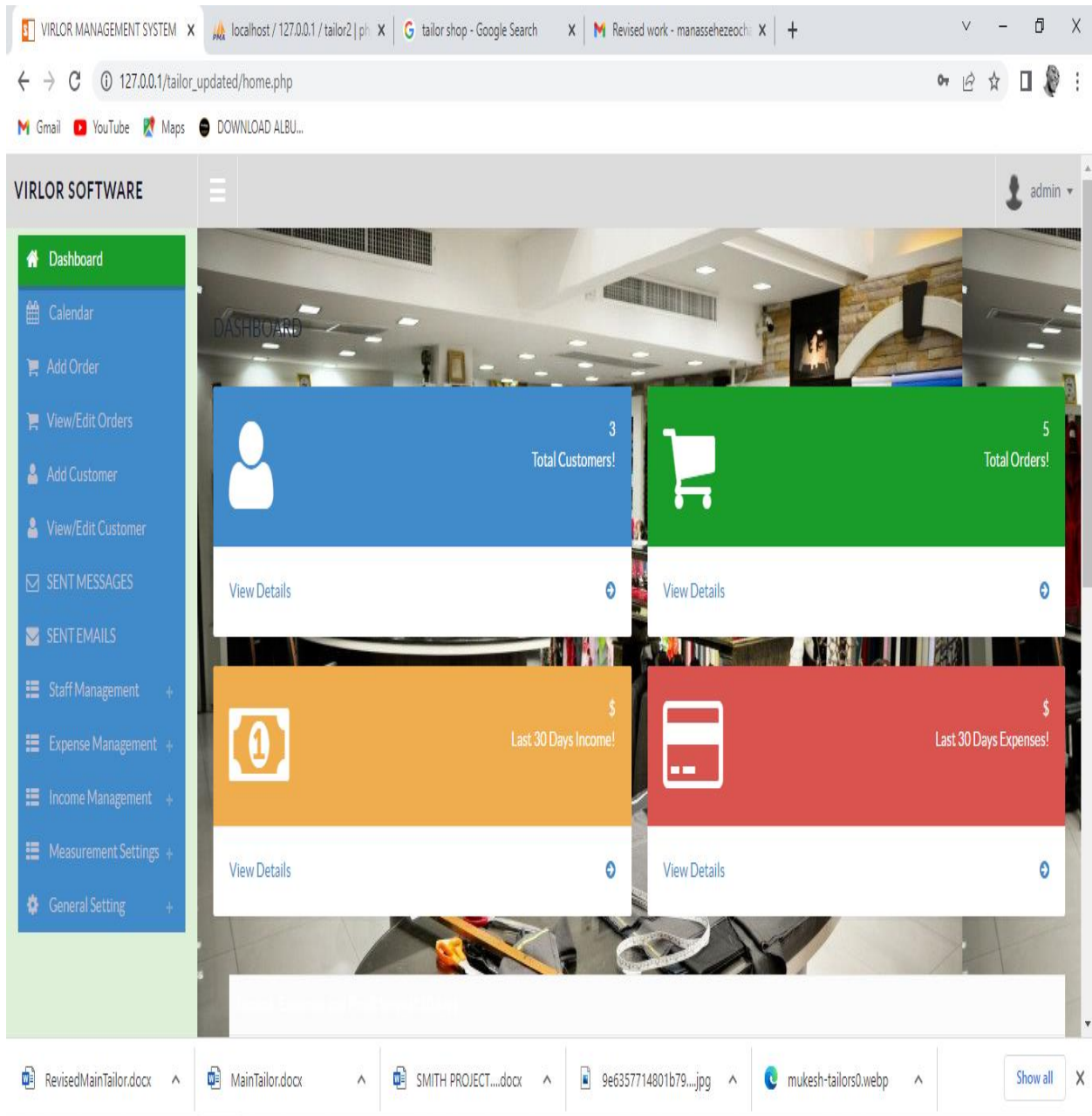
The figure above shows a screenshot of the systems homepage, it is a system non-CLUSTERED page which comprises of a SIGN IN/LOG IN button which makes it easier for users to operate, after user details imputation, the system verifies login details and grants access or denies.

4.4 Implementation of the subsystem

This shows all the activities performed by the admin during the log in process.

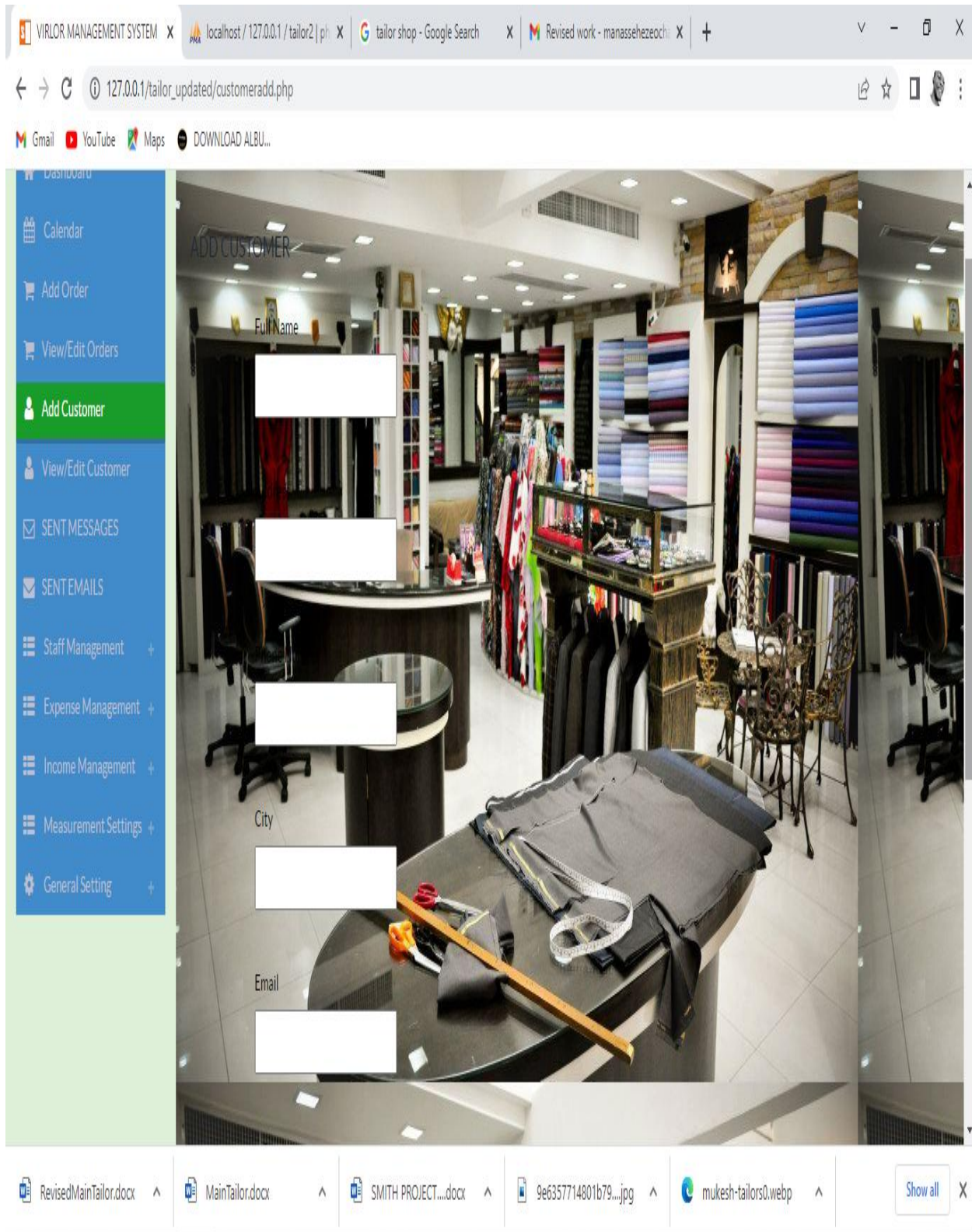


In the figure above, the admin logs in to provide the system with up-to-date information and to view the operations as the back end of the system (i.e., check in and check out user operation). The admin has the sole objective of feeding the system with up-to-date information and upgrading it time to time. The admin cannot create users but only put up more information for the users.

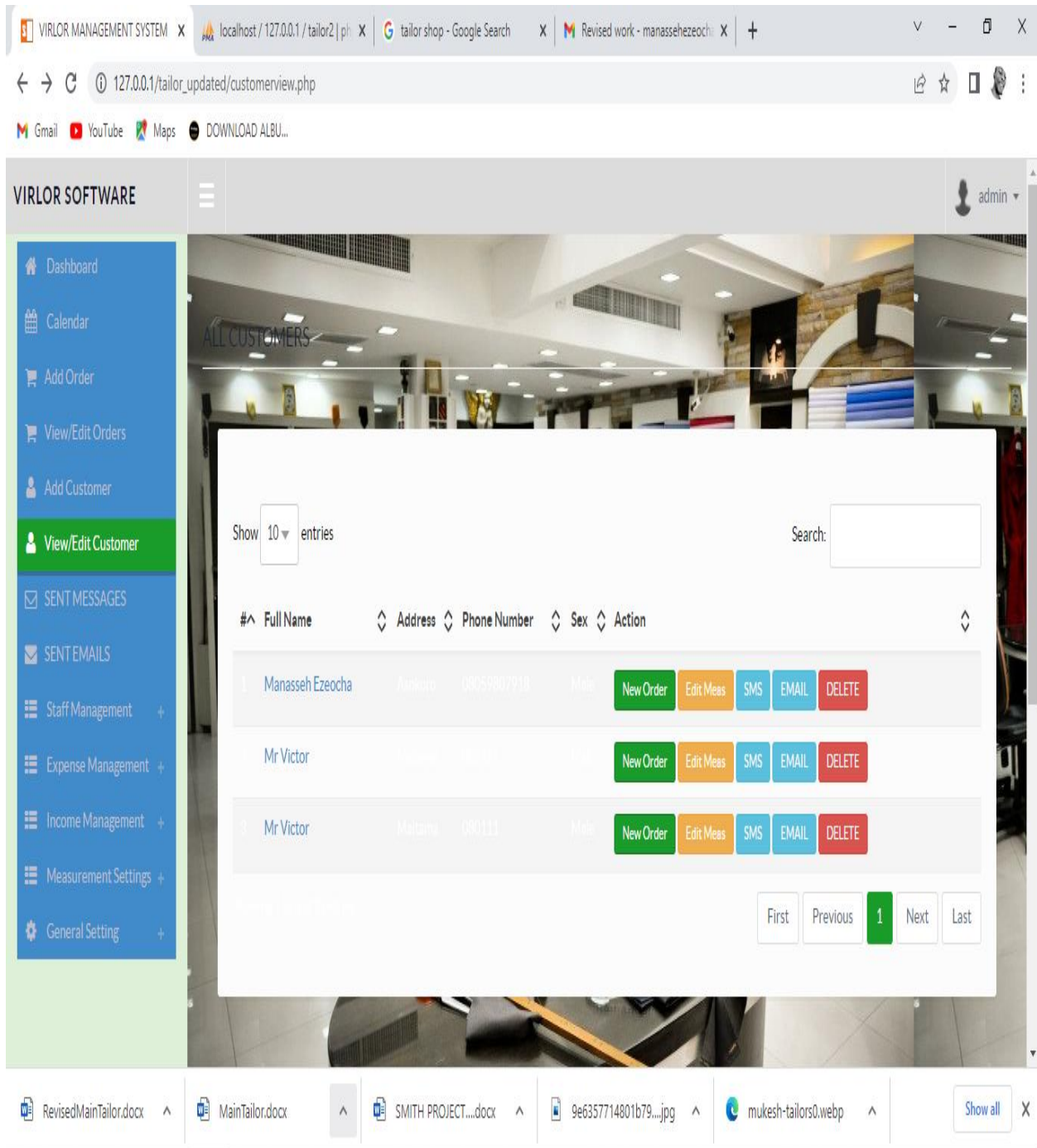


4.5 Query subsystem implementation

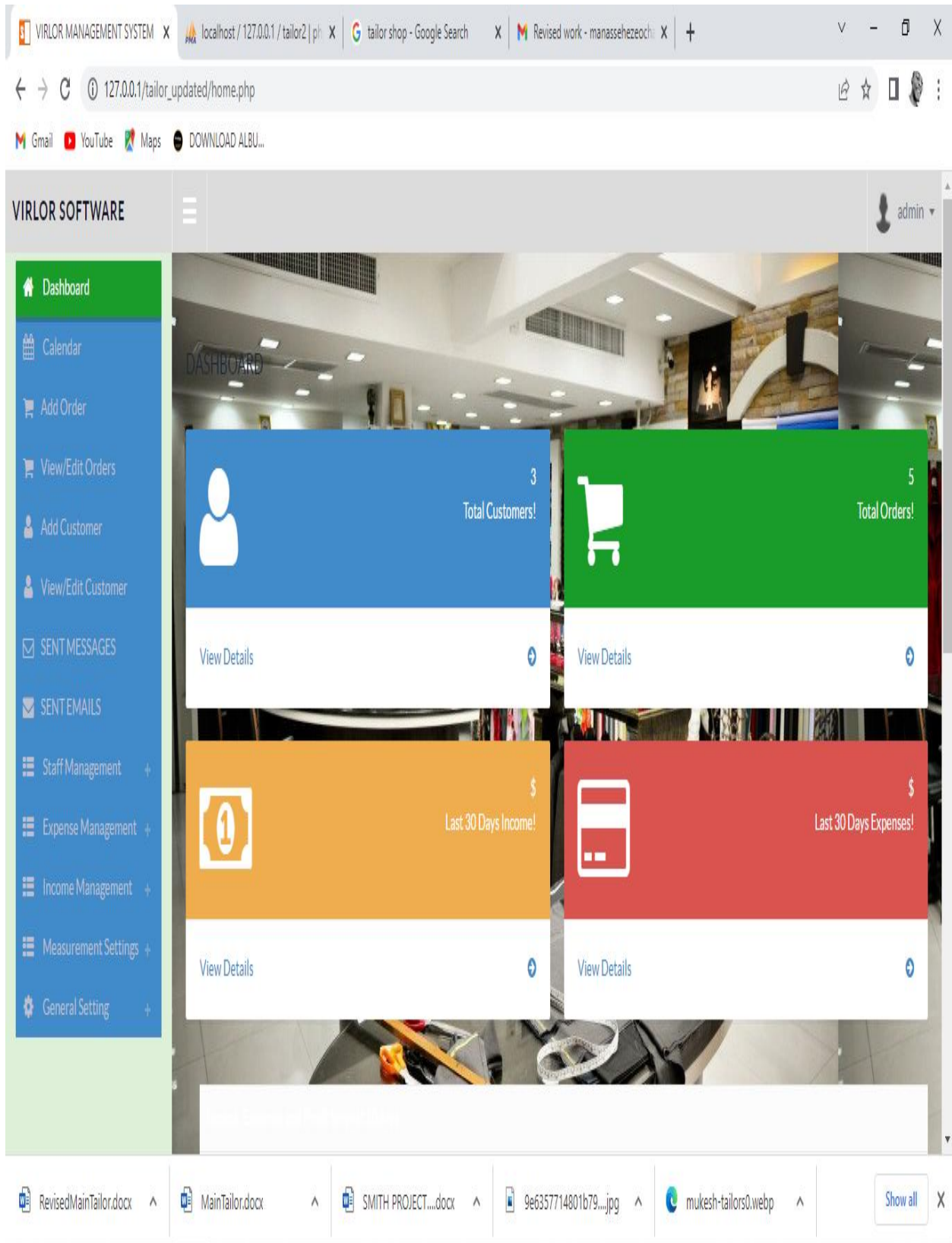
The figures above show the subsystems. The query subsystem explains how the users interact with the system and how the system responds to users. In other words, it shows the communication process between a system and its user. It shows the responses users get when they go through the different sections of the system.



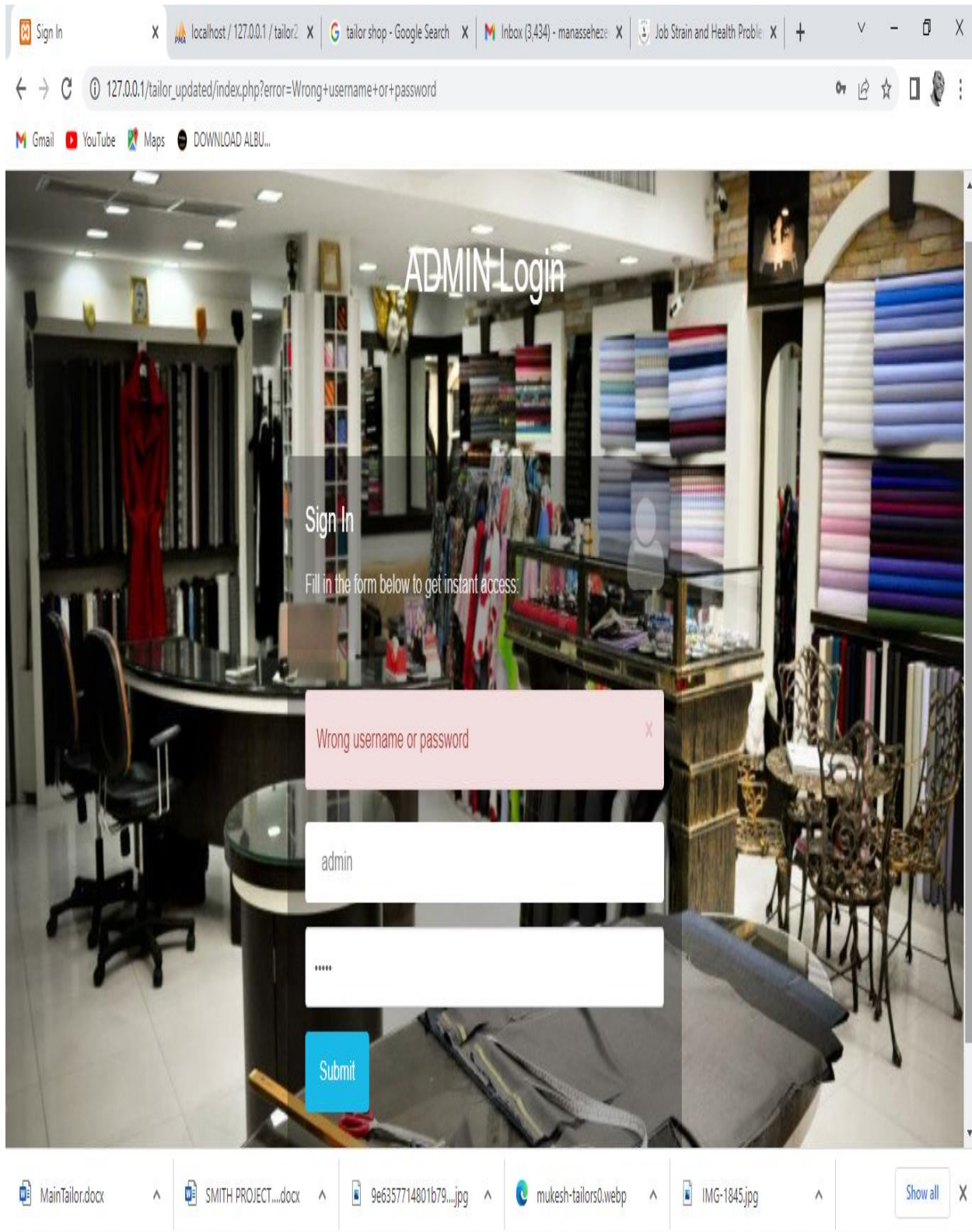
This section displays the add in form presented to new and existing users of the system; it is a non-CLUSTARD page which enables tailors carry out this function without further navigation.



The figure above shows the form for users wish to edit details earlier inputted into the system. This form allows users to create their account so they can have access to the system and every information they need.

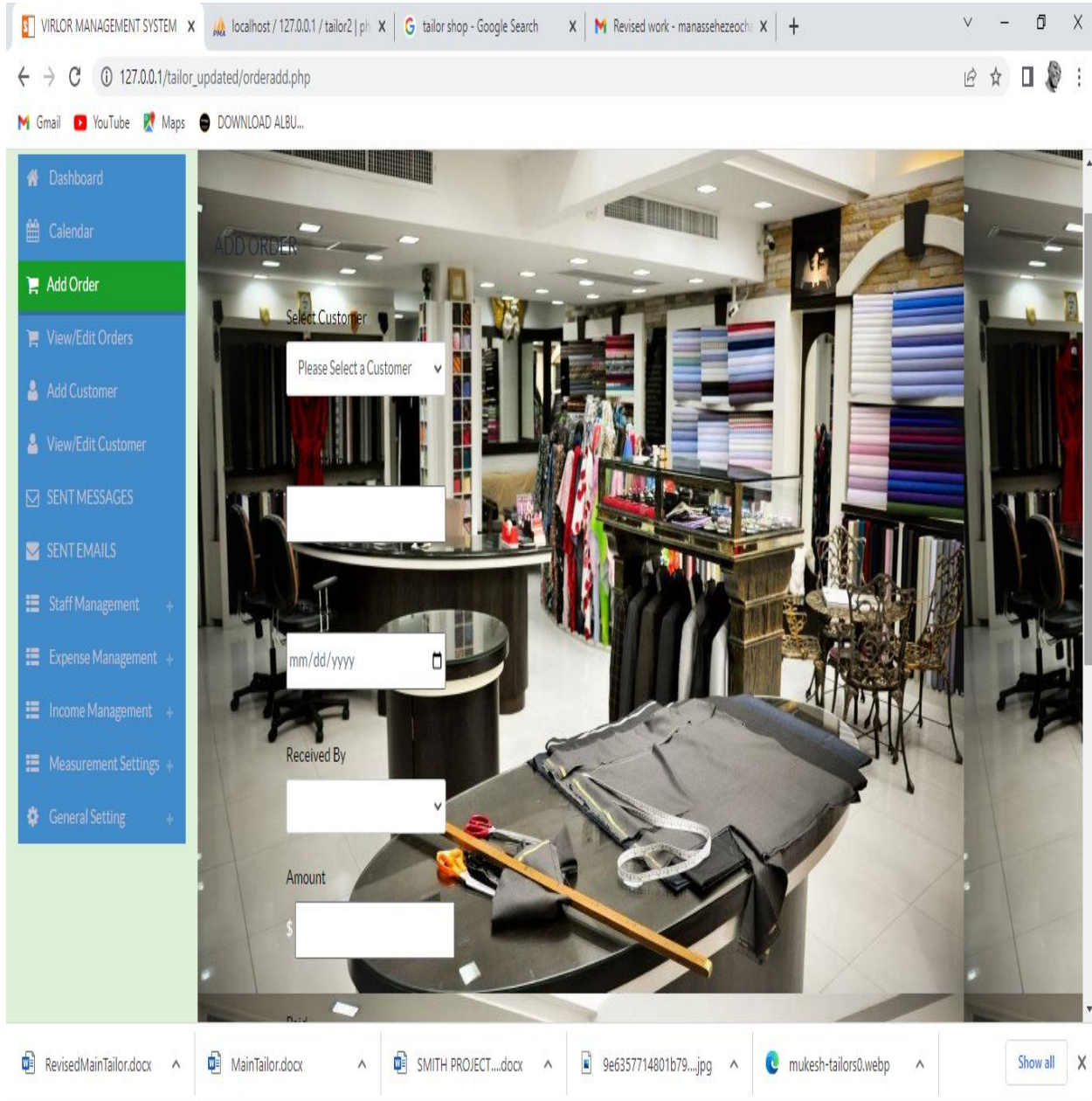


The page above shows the display users first see upon log in. it consists of different buttons such as user profile.



The figure above represents an Error. This is as a result of wrong login details, if a user inputs wrong username or password, the system would give an alert that an error has occurred. This can also happen when a user does not exist on the system and he goes straight to login in

instead of signing up first. The system doesn't recognize such details in the database hence it denies the user any access.



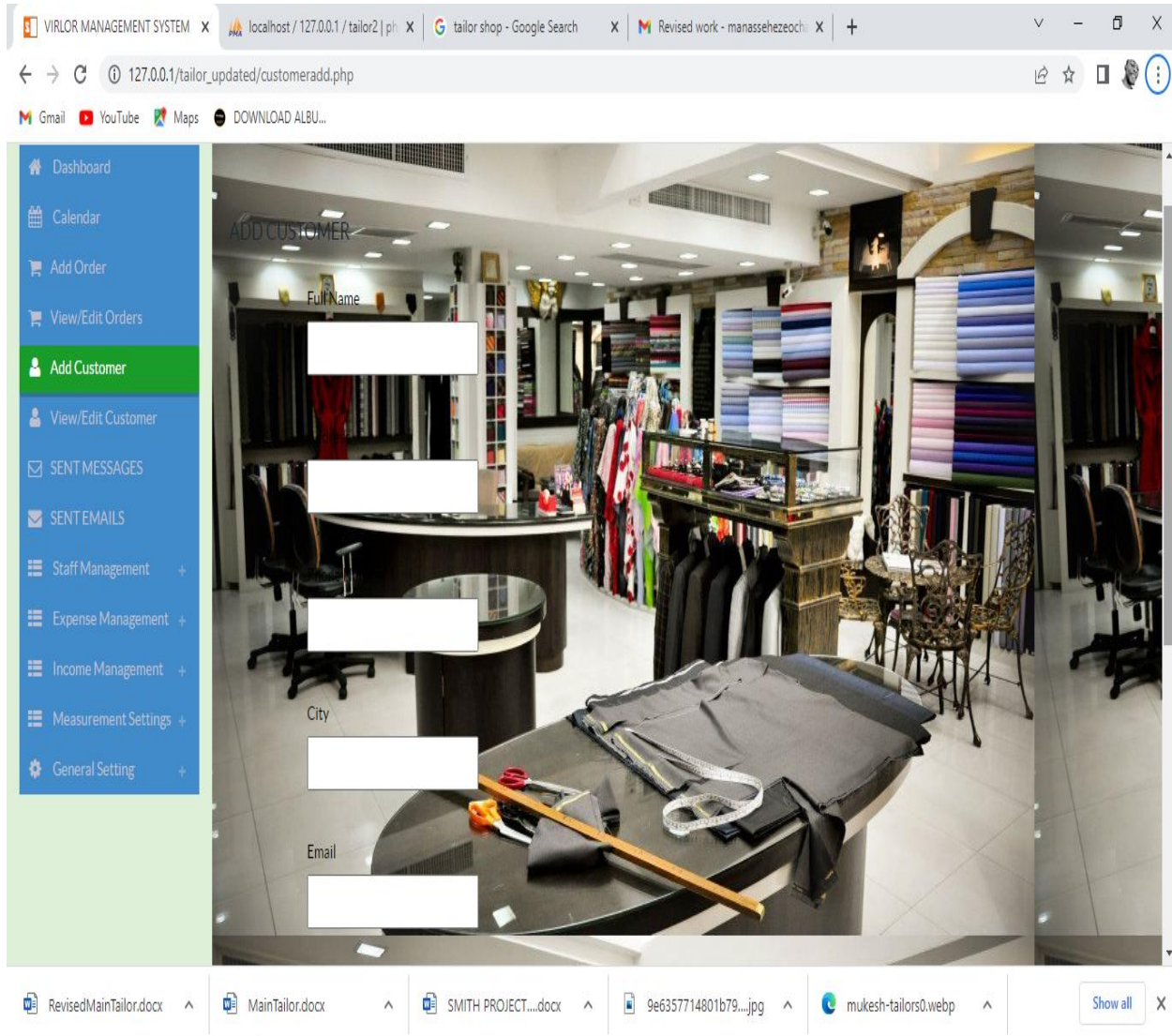
The figure above represents the booking form with which users are able to make their reservation for a particular tailor, and every detail asked of the user is necessary to make a reservation. If any field of this booking form is left blank, user would be unable to make a tailor reservation until fields are completely filled.

4.5.1 System functionality process

Virilor system works in a systematic order, meaning you cannot skip step 1 and head into step 2, i.e., a particular order. The system functions are as follows:

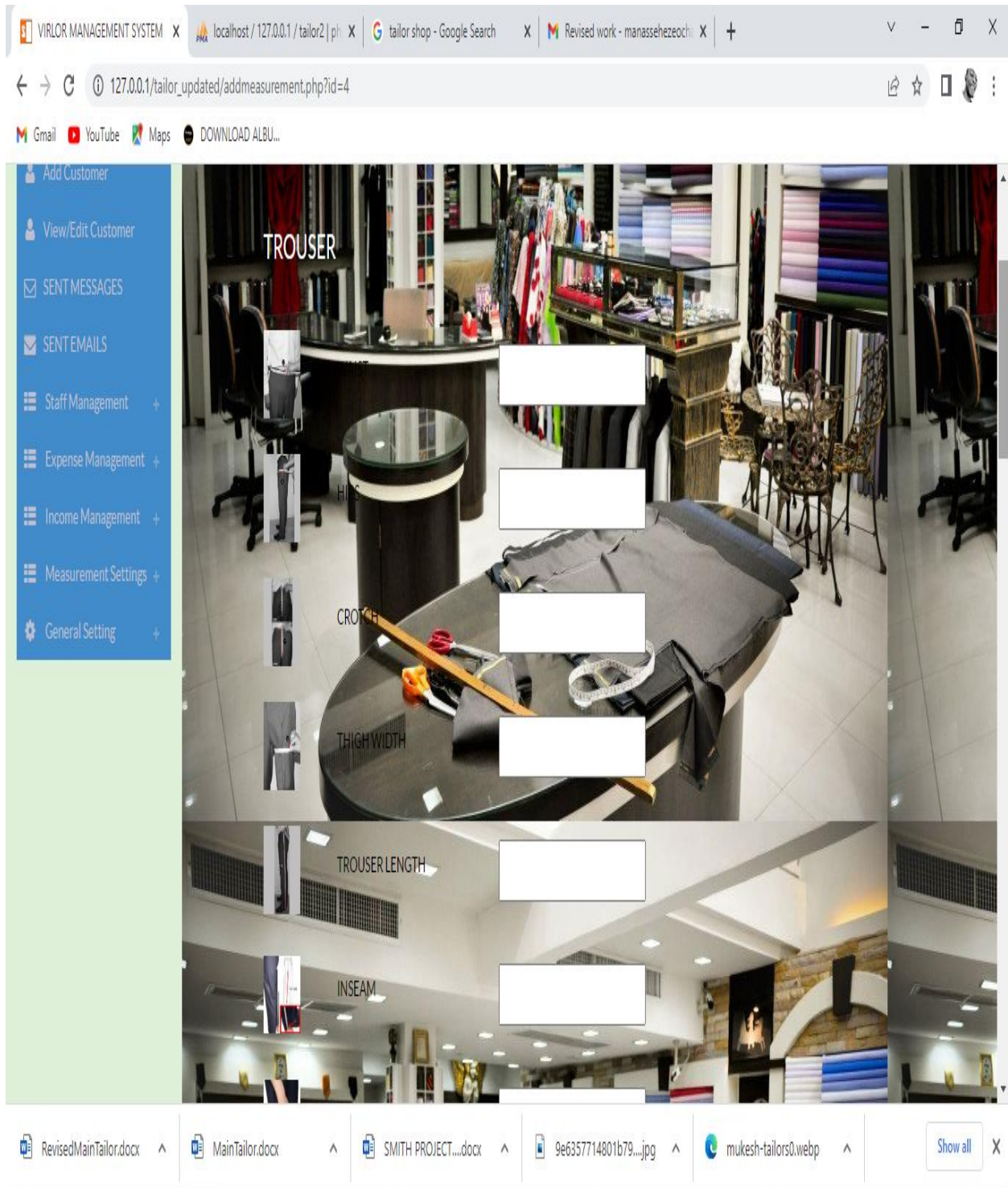
- i. Adding of customers;

This is the first step in the tailoring process, a customer must be added into the system first. After which a form is displayed which request customer information such as, email address, house address, city, phone number and full name.



Once this form is filled, the system directs tailor to the next field, which is
i. Measurement submission

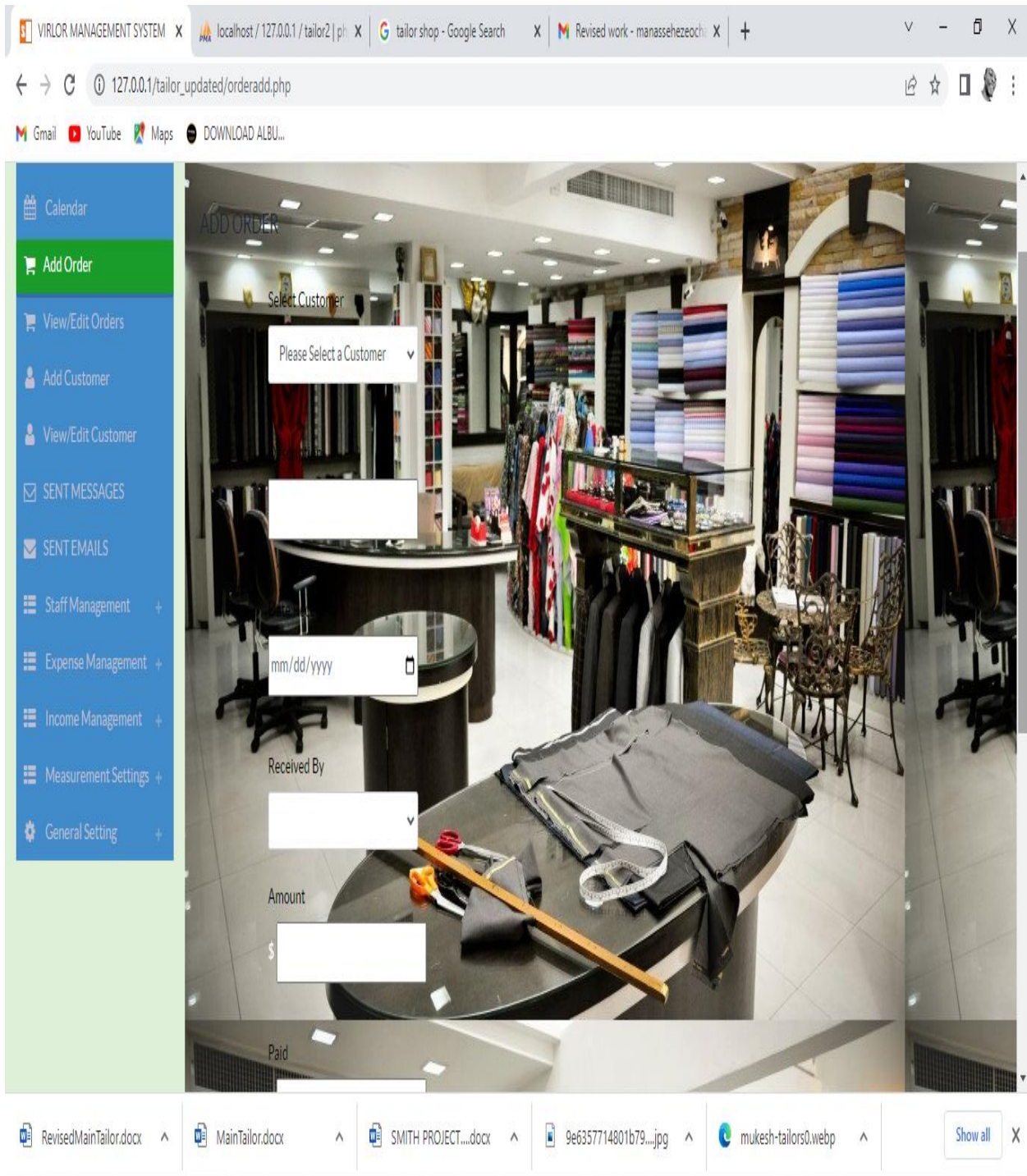
Here the measurement needed to sew the customers attire is requested.



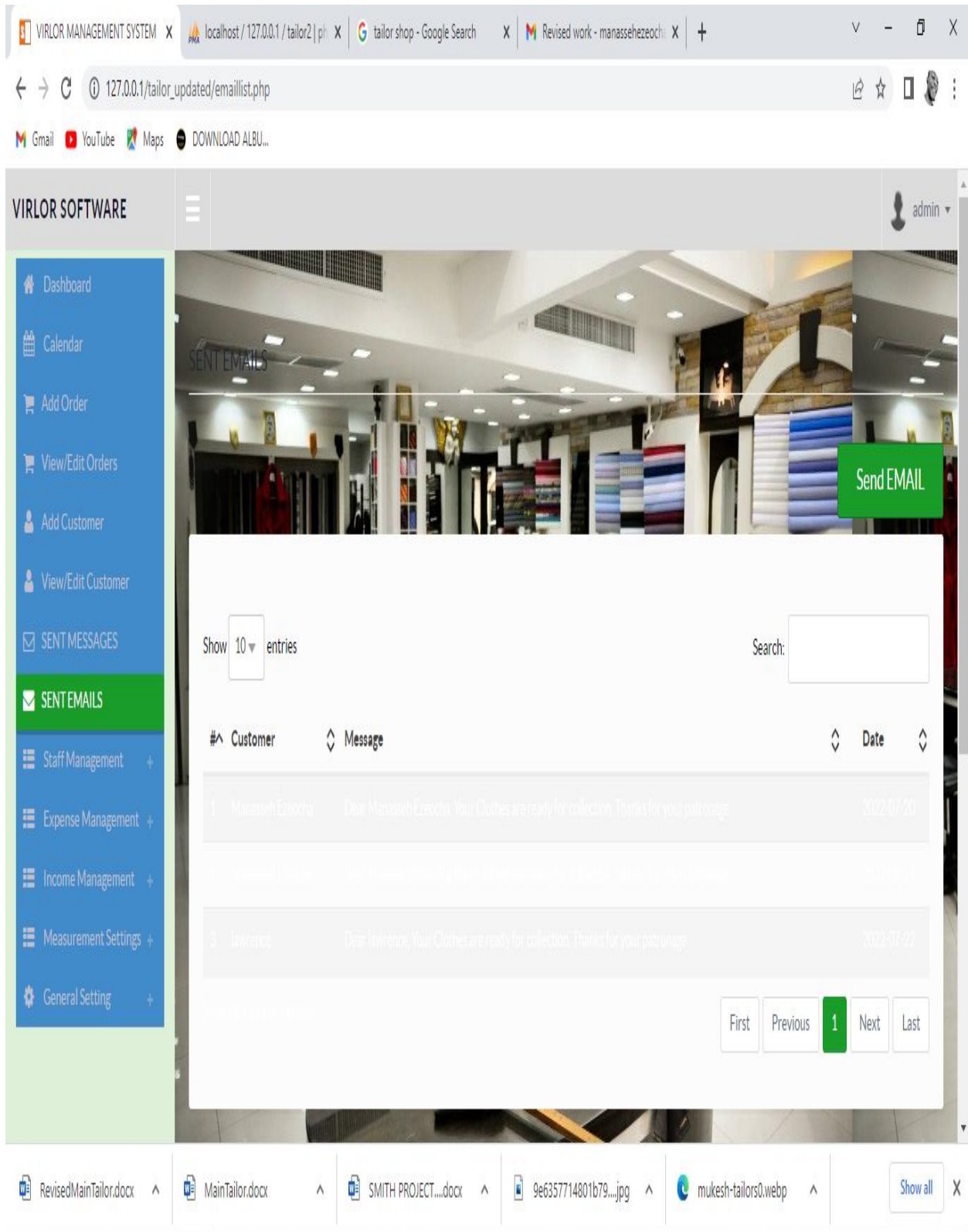
Once the measurement form is filled, the system directs the tailor to the next form, which is

i. Add Order

Here the customer order is recorded and once complete, Virilor system sends a message to the customer, informing them that their attire is ready.



Once order is completed a message is sent to customer email indicating that their clot is ready.



4.5.2 System testing and integration

Although this system can serve as a stand-alone system, it can also be used to complement the existing Manuel system. This kind of implementation used in this project can be best be described as Parallel implementation, because it can be used alongside the already existing Manuel system.

This system was tested by 5 customers who registered on the system, with a particular tailor, made bookings and got evidence of service via receipt. This system would ease the stress of the Manuel tailoring system already in existence. It would provide accurate, and efficient output.

The system will save people who are in further geographical locations the stress of coming to pass through the Manuel system of making enquiries, going to the tailor store, registering measurement on a sheet of paper, and expending excessively on transportation.

5. Summary and conclusion

This chapter describes and summarizes the objective of the system stimulated above, the limitation of my study, conclusion and recommendation.

5.1 Introduction

This project work 'design and implementation of a virtual tailoring management system' is a revolutionary approach to improve the management of the tailoring service it conventional and Manuel practices. it designed a path however in which the existing Manuel system of handling records and books can be deployed through a dynamic system and would enable client choose the particular one (tailor service) they desire to visit.

5.2 Summary

This research work gave a break down on the concepts of tailoring, what tailoring is really about, and the real essence of a virtual management system. It has in every way achieved the objectives which led to the genesis of this project. In this study, the aforementioned objectives which are: to identify and explore the challenges experienced in the management of existing Manuel system, and accurate, design a system that will make delivery of services more effective and efficient.

The chapter two of this work clearly showed the review of other people's works, the design and implementation of this proposed project. Hence chapter three gave a breakdown analysis of the existing Manuel system, its problems.

four on the other hand talked about the complete implementation of the designed system, the tools used to implement this system and their functions.

5.3 Findings of discussion

In the course of achieving the aim and objectives of this project, the existing Manuel system was studied in detail and all its problems or hindrances were noted, and a room for improvement was generated.

I figured that with the world digitally globalizing, developing a tailoring service management system would be key. It also gives the tailoring profession a chance to show the world the true beauty of its practices and methodology.

5.4 Suggestion for future

This project is limited due to time restrictions and lack of resources. the scope of this study could therefore be broadened with enough time and good allocation of resources. it is here recommended that this research project be thought through, and improved upon. It will take the tailoring workers association to a greater step ahead in technology instead of the archaic method of management.

It would be very useful not just for Abuja but also for many other states and country at large.

5.5 Limitation of study

The Virlor system modernizes most aspect of the tailoring service except taking of measurements, clients who can't take measurements themselves may be encountered along the line, clients who don't have measurement materials can be encountered as well, this is a major issue that is faced and a limitation to my system.

5.6 Conclusion

In this project, we have been able to understand and explore the problems faced by existing Manuel system, design a tailoring services management system to tackle these problems, and implement an application that will help people understand the tailoring process properly, appreciate it and promote it. The success of any organization depends on effective and efficient management. Hence the success of the tailoring service management system.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

References

- Cletus, I. (2020). *Online Tailoring Management System*.
<https://www.graciousnaija.com/2021/11/online-tailoring-management-system.html>.
- Cooper, H. (1998). *Synthesizing Research: A Guide for literature Reviews*.
https://books.google.com/books/about/Synthesizing_Research.html?id=gAxHAAAAMAAJ&source=kp_book_description.
- Custom tailors and designers' association (2019). *Opening a Tailor Shop*.
<http://www.gaebler.com/Opening-a-TailorShop.htm>.
- Hardy, V., & Hauge, J. (2019). Labour challenges in Ethiopia's textile and leather industries: No voice, no loyalty, no exit? *African Affairs*, 118(473), 712-736, <https://doi.org/10.1093/afraf/adz001>
- Mathew, A. (January 2022). *Alison Mathew's "Tailoring"*. <https://fashion-history.lovetoknow.com/fashion-clothing-industry/tailoring>.

- Mutembei, D. (October 13, 2021). *Online Tailoring Management System*.
https://www.academia.edu/12426471/ONLINE_TAILORING_MANAGEMENT_SYSTEM.
- Gieves & Hawkes (n.d.). *No. 1 Savile Row*. <http://en.wikipedia.org/wiki/Bespoke>
<http://www.bbc.co.uk/britishstylegenius/content/21811.shtml>.
- Idrees, S., Vignali, G., & Gill, S. (2020). Technological advancement in fashion online retailing: a comparative study of Pakistan and UK fashion e-commerce. *International Journal of Economics and Management Engineering*, 14(4), 313-328.
- Kolade, C. (August 30, 2021). "What is PHP? The PHP Programming Language Meaning Explained".
<https://www.freecodecamp.org/news/what-is-php-the-php-programming-language-meaning-explained/#:~:text=What%20Does%20PHP%20Mean%3F,recursive%20acronym%20for%20Hypertext%20Preprocessor>.
- Lancaster (2013). *Tailors in the UK*. <https://www.stears.co/article/the-state-of-nigerias-fashion-industry>.
- Lerdorf, R. (June 8, 1995). *Announce: Personal Home Page Tools (PHP tools)*.
<https://groups.google.com/g/comp.infosystems.www.authoring.cgi/c/PyJ25gZ6z7A/m/M9FkTUVDfcwJ?pli=1>.
- Ogunfuyi, K (2019). The state of Nigeria's fashion industry. Retrieved from Lancaster (2013) Tailors in the UK. <https://www.stears.co/article/the-state-of-nigerias-fashion-industry>.
- Ominijei, E. (2019). The role of institutions in promoting entrepreneurship in the Nigerian fashion industry (dissertation). Retrieved from <http://urn.kb.se/resolve?urn=urn:nbn:se:sh:diva-39330>.
- Oswald Spring, Ú. (2019). *Development of underdevelopment*. Úrsula Oswald Spring: Pioneer on Gender, Peace, Development, Environment, Food and Water: With a Foreword by Birgit Dechmann, 268-284.
- Priya Dwivedi, & Kiran, U. V. (2020). *Job Strain and Health Problems among Tailors*.
<https://www.ijsr.net/archive/v4i8/SUB157254.pdf>.
- Resources For Entrepreneurs (Start Your Own Business) (July, 2022).
<https://www.gaebler.com/Starting-Your-Own-Business.htm>.
- Rudestam, K. E., & Newton, R. R. (2020). *Comprehensive guide to content surviving your dissertation process*. Newbury Park, CA: SAGE.
- Sandell, L. (2017). Henry Poole & Co.: How a 200-year-old bespoke tailor have managed to stay modern. (Dissertation). Retrieved from <http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-12839>.
- Shaw, G. (2019). *Tailored clothes*. <http://www.askmen.com/fashion/keywords/tailored-clothes.html>.
- Sitlong, N. I., Nonyelum, F., & Ogwuleka (2020). Harmonization of tools and techniques for system development. *World Journal of Innovative Research (WJIR)*, ISSN: 2454-8236, 9(1), 106-114.
- Stegmaier, M. (2019). "If you're not already writing, stop!": An Interview with Michael Tolkin. *Zeitschrift für Anglistik und Amerikanistik*, 67(4), 429-441.
- Waugh, J. (2001). "Give This Man Work!": Josephine Shaw Lowell, the Charity Organization Society of the City of New York, and the Depression of 1893. *Social Science History*, 25(2), 217-246.
- Yourdictionary (2019). *What is a tailor and what do they do*. <https://www.yourdictionary.com/tailor>.





Automated Assessment System Using Machine Learning Libraries

Victor Adebola Omopariola, Chukwudi Nnanna Ogbonna & Felix Uloko

*Veritas University, Abuja, NIGERIA
Faculty of Natural and Applied Science*

Monday J. Abdullahi

*Airforce Institute of Technology, Kaduna, NIGERIA
Department of Computer Science*

Received: 7 January 2023 ▪ Revised: 23 July 2023 ▪ Accepted: 25 October 2023

Abstract

Assessment and the grading of students is a task that has been done for as long as school has existed. This was previously done by teachers in primary and secondary, lecturers for institutions like JAMB and lecturers in schools. Up until now students' marks were influenced by other external factors such as bad handwriting, lengthy paragraphs, roundabout way of speaking rather than going straight to the point and the sheer number of assignments the lecturer has to mark. This has resulted in students getting lower or higher marks than they should be awarded. This project is to create an ML (Machine Learning) powered assessment system that will take the assignment questions and the marking scheme and award the student the marks similar to what the ideal lecturer would have given. This will also reduce the time the lecturers spend on marking and ensure the students get their results on time. This project will be made with Python and machine learning and will be tested with a number of potential answers to the questions and their grading's. This will enable system to be able to grade assignments as soon as they are uploaded. This research will be limited by the fact that the system can only handle the marking of short sentences accurately and not long paragraphs. The system is also limited by the fact that it can only mark with the aid of the marking scheme and not without it so it is not a truly intelligent model in that regard. The research showed that the system is indeed capable of obtaining the similarity between two paragraphed answers provided but it needs extras to produce the most accurate results.

Keywords: assessment, automated assessment system, machine learning library.

1. Introduction

1.1 *Background*

Assessments generally refer to the tools educational overseers or educators use to evaluate, measure or determine the educational capacity of a student, the readiness of a student to learn what has been taught to him/her over time and the educational needs of the student being taught. According to Stassen et al., assessment has defined as "The systematic collection and analysis of information to improve student learning." As humans' assessments is very important

© **Authors.** Terms and conditions of Creative Commons Attribution 4.0 International (CC BY 4.0) apply.

Correspondence: Adebola Victor Omopariola, Veritas University, Department of Computer and Information Technology, Abuja, NIGERIA.

for tracking progress for all ages. Generally, this has been done through the use of assessments created by various educational bodies and the various standards set. There are different tests that directly correlate to how much a student has learned about the concepts or information the educator is trying to teach them. There are various testing methods which have been developed throughout the years to determine the level of knowledge and creativity of students. Examples of these testing methods include peer-assessments; this is when a student marks another's assessment using a set of rules or guidelines this helps to improve a student's judgement and learn the processes to a result being awarded. Presentations which usually involve the student delivering a piece in front of either a class or assessors and they get graded for it, discussions mostly in the form of a debate to gauge the knowledge of the students especially when their perspective is being challenged, reports and the most common written assessments; these can come in the form of time constrained individual assessments and these have the unintended effect of surface learning and cramming and can come in various forms such as open-book, in-class assessment or take-home assessments. Generally, assessments cover a range of subjects which include; Expressive arts, Health and wellbeing, Languages (including English, French, classical languages and modern languages), Mathematics, Religious and moral education, Sciences, Social studies, Technologies. Assessments are done periodically and at specified intervals set by either the school or assessment body, such time could be at the end of a school year or at the end of a semester but one thing they all have in common is that they occur at key points in a student's learning journey. Assessment is usually split up into two categories or two purposes (these differ from the method the assessments are carried out in), specifically summative assessment and formative assessment. Summative assessments are done at the end of a course and as a result it "sums up" everything the student has learned from the beginning of the course to the end of the course. They are generally done with the use of comprehensive final assignments or papers. The second type is formative assessment which is done during the students learning time this kind of assessment is done in order to enhance the learning experience of the student. This kind of assessments are done for the sole purpose of sharing the results back to the students so they can understand their strengths and weaknesses and reflect on them. This type of assessment typically includes things like coursework and others. Grading is not supposed to be the major aspect of assessments as assessments is supposed to make sure the students attain the knowledge required of the course. Also Grading does not tell you about the students individual learning outcomes that have been achieved. Grades are now one of the most important parts of any schooling although they may not accurately reflect the level of the student's skill or understanding. The job market revolves around how well you did in university or college with top grades such as "First Class" or "Second Class Upper" being preferred across universities regardless of the rank of the university the grade is obtained from. According to Allen et al. (2001), "the larger the variability in grading practices from teacher to teacher and from school to school, the more limited the value grades have as guides for planning the academic and career futures of students."

1.2 Problem statement

From the problem statement above, we see how crucial the grading process is to the future of students and more often than not scripts are graded differently by different teachers in the same discipline. This project aims to increase ease of marking assignments and CA's and provide the students with an indisputable result and also to increase the confidence of the students when writing assignments and the performance of the students over time.

1.3 *Research questions*

1. How efficiently can a system grade a student's assignment?
2. How accurate is the grading of a student's assignment by an artificial body?
3. The influence of computer assisted grading on the morale of the students and the education system.
4. The time taken to deliver assignments after they have been graded automatically by the system.

1.4 *Research aim and objectives*

- To investigate whether the involvement of a system to mark and grade students provides a true representation of their skill and increases their confidence in writing assignments.
- To determine if a system is capable of marking assignments using current existing algorithms.
- To investigate if a paragraph styled answers can be appropriately compared without error.
- To find out if existing algorithms can compare answers of varying lengths.

1.5 *Research motivation*

The motivation for this research is based on the fact that many factors play a role in the grading of a student's paper such as handwriting, lecturer mood and the fact that assessments can be subjective depending on whether the lecturer is strict or lenient. This all affects a student's final grade and result so through this project I am attempting to solve some of the issues with the assessment process leading to fairer results and increased satisfaction with the schooling system.

1.6 *Significance of the research*

This research can be applied by all various universities, secondary schools and even the ministry of education to remove disparities when the assessment system and increase the confidence of the students in the system.

1.7 *Delimitation of the research*

This research will be limited by the fact that the system can only handle the marking of short sentences accurately and not long paragraphs. The system is also limited by the fact that it can only mark with the aid of the marking scheme and not without it so it is not a truly intelligent model in that regard. The system will be limited by the fact that it can only compare answers in English and not in other languages or disciplines. Another limitation is the fact that the documents have to only contain the answers and no other additions such as a cover page for the system to work effectively.

2. Literature review

Assessment is an integral part of the learning process and an accurate method of gauging how much a student has learned throughout the duration of the course. This process has however not been without issues. As I stated in the background of this study, according to Allen et al. (2001), “the larger the variability in grading practices from teacher to teacher and from school to school, the more limited the value grades have as guides for planning the academic and career futures of students.” More importantly validity and reliability of grading practices used in the marking process have had a profound effect on the futures of students Allen et al. (2001) also said in their paper that “Since important decisions are often based on a student’s grade, unreliable and invalid grades may result in dire consequences for the student. Invalid grades that communicate an understatement of the student’s understanding may prevent a student with ability to pursue certain educational or career opportunities.” He even goes further to state that there could be consequences of giving a student more marks than he/she deserves this could lead to the student, after graduation being put in situations that his/her grade says they are fit for but in reality they are underprepared or they have an inadequate level of information required for the role or position. Also supporting my theory that the grading system of schools does not take account the status of the school he states “Research indicates that when compared to schools in more affluent areas, students in low SES schools receive grades that are two letter grades better than students in affluent schools when national standardized scores are held constant.”

Cizek (1996) in his paper states that one of the crucial aspects of grading that needs to be touched is the training of teachers in grading practices based on sound measurement principles relevant to their classroom lives. In the grading process it is shown that the classroom actually affects a final grade a student gets in addition to assignments. Cross et al. (1996) stated that “Some studies have found that 2 out of 3 teachers believe that effort and student conduct and attitude should influence final grades of students.” This goes to show that while a student may have outstanding performances in written assessments if they do not show substantial effort in their classes the lecturers are unlikely to award them the marks they deserve even if they show the adequate amount of knowledge required to earn the grade. Also, regarding the problem that the grading process is subjective. Allen et al. (2001) quote that “All grading is at some level inherently subjective. However, teachers need to recognize the subjective factors in order to reduce them as much as possible to increase the objectivity and validity of their assessment and grading practices.” This could be due to problems such as student handwriting, student behavior, the number of scripts given to the teachers to assess and others. Regardless of the method taken to reduce subjectivity it still won’t change that fact. This is why the proposal of a system to mark scripts automatically and fairly is an important one. Systems won’t be affected by the problems suffered by the lecturer or won’t have prior experience with the student outside the questions and answers provided to it.

Though this project aims to remove bias from the marking of assignments it doesn’t affect the in-class scores of the students during assignments. For example, if a faculty were awarding 30 marks to classwork’s and tests, 40 marks to in class contributions and participation while the rest 30 marks is gotten from the assignments, this will ensure a student gets the best marks from the classwork’s and the assignments whereas the other 40 is within the direct control and discretion of the lecturer or teacher. This would also correlate with Allen et al. (2001) study which showed that “This would seem to imply that a grade is used to communicate not only how much content knowledge one has achieved, but also how well one has complied with the teacher's requirements.” In my opinion this helps to stress that fact that although assignment grading alone is not enough to give a student an excellent result it should not be overlooked as in a scenario where a grade was awarded perfectly then the rest is up to other aspects of the student which is an ideal situation.

In a paper by Tomkinson et al. (2011), they stated that a strategy taken to prevent the incorrect marking of student assessments specifically in the undergraduate sector is to have multiple markers and then use the average as the actual result for the student. Tomkinson et al. (2011) also talks of a “halo” and “horns” effect where the halo effect refers to where “supervisors give higher marks than the written work merits because they have been aware of the effort and thought processes” and the horns effect is when “a dilatory student produces a dissertation of greater merit than the supervisor has been led to expect.” Also, another problem pointed out by Tomkinson et al. (2011) with using second or third markers is that they may not have sufficient knowledge concerning the subject area and as such cannot mark the student to the degree of accuracy expected. These are the kinds of problems which would be solved by the introduction of my system as there won’t be a need for a second or third evaluator.

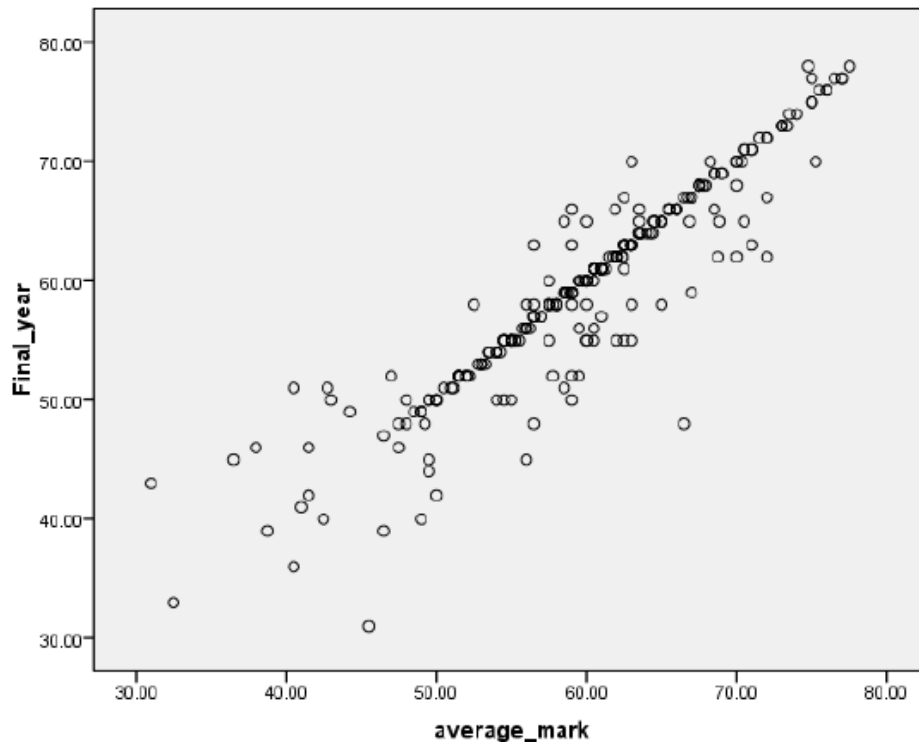


Figure 1. Final marks against average of first and second markers

The results of Tomkinson et al. (2011) research showed that the second markers who are generally not the lecturers of the course award lower marks on average while the first markers award higher but when the average has been collected it shows consistency with each of their markings.

AI which is short for Artificial Intelligence and it mainly deals with creating systems that can imitate intelligent human functions. In particular NLP is one of the most challenging aspects of AI due to the fourfold nature of it; Speech Recognition, Syntactic analysis, Information Extraction and Discourse Analysis. It mainly deals with human language which can have different meanings stemming from the same sentence due to differences in things like punctuation and two completely different sentences can have the same meaning. Artificial Intelligence can be incorporated into the learning process and also make it more efficient and better. Luckin (2017) stated in her article that “AI is a powerful tool to open up the ‘black box’ of learning.” AI has already been employed in the assessment of essays known as AES (Automated Essay Scoring) with the most popular of this category being IntelliMetric. These systems have high reliability and are able

to grade and provide feedback within seconds something teachers cannot do. This category of software is unique in the sense that it wants to understand the meaning of the text and as such determine if it relates to the question asked and grade it appropriately.

The system I propose is one that has a set of questions and a sample answer to the questions set by the educator and the system will perform a comparison between them to determine their degree of similarity. This removes the need for the system to fully understand the text and the hidden contexts but it does take depth away from the system. The proposed system will also have the added bonus of evaluating if the educator actually has an understanding of the assignment or assessment set as they will have to provide their own answers to the question. This way it will remove inherent issues like the one Dikili (2006) stated in her paper regarding PEG (Project Essay Grader); “Since PEG used indirect measures of writing skill, it was possible to trick the system, i.e., writing longer essays.”

Although this would be an optimal solution to solve some of the multiple issues with traditional grading it does come with its own drawbacks. One of the issues is what (Luckin, 2017) highlighted in her article; she states AI can be very costly to implement especially when looking at massive systems that can handle the grading of multiple students across multiple subjects. Also, to complete a system that could have such capabilities it would need to be backed by some national or large corporation as well as have access to the core details of the student and the curriculum as well as being able to determine which areas the student is struggling with for future development. Luckin (2017) stated “this suggests an annual budget of US\$600 million per year for a complex AI project. It therefore seems reasonable to suggest that a country, such as England, might need to spend the equivalent of US\$600 million (£500 million) per year to make AI assessment a reality for a set of core subjects and skills.”

Previously the only work done on using systems for assessments was in the area of multiple-choice questions and similar shown by researches by Ana et al. (2013) and Boussakuk et al. (2021). But Wilson et al. (2000) proposed the creation of the system BEAR. This stands for Berkeley Evaluation and Assessment Research the proposed system would be capable of understanding the curriculum of the school and in turn accessing the students throughout the year and through their work. They used the IEY (Issues, Evidence and You) developed by SEPUP (Science Education for Public Understanding Project) Course to test their program. The program had a grading scheme which was adapted from SOLO Taxonomy by Biggs et al. (1982). According to Wilson et al. (2000), their grading scheme is a system ranging from 1 which is “an answer with only one correct aspect to it” to 4 which is a perfect answer by the student.

In Wilson et al. (2000) paper they also express the need conform to standards of fairness which they state as including “Consistency and Unbiasedness.” This also correlates with some of the main problems with traditional grading that I outlined in the earlier parts of the literature review. Ultimately this study was just for a system to evaluate the grading by teachers and where the students need help as opposed to a standalone AI that can grade the students by itself and also provide feedback on where the student got it wrong.

Evidence and Tradeoffs (ET) Variable

Score	<i>Using Evidence:</i> Response uses objective reason(s) based on relevant evidence to support choice.	<i>Using Evidence to Make Tradeoffs:</i> Response recognizes multiple perspectives of issue and explains each perspective using objective reasons, supported by evidence, in order to make choice.
4	Response accomplishes Level 3 AND goes beyond in some significant way, such as questioning or justifying the source, validity, and/or quantity of evidence.	Response accomplishes Level 3 AND goes beyond in some significant way, such as suggesting additional evidence beyond the activity that would further influence choices in specific ways, OR questioning the source, validity, and/or quantity of evidence & explaining how it influences choice.
3	Response provides major objective reasons AND supports each with relevant & accurate evidence.	Response discusses <u>at least two</u> perspectives of issue AND provides objective reasons, supported by relevant & accurate evidence, for each perspective.
2	Response provides <u>some</u> objective reasons AND some supporting evidence, BUT at least one reason is missing and/or part of the evidence is incomplete.	Response states at least one perspective of issue AND provides some objective reasons using some relevant evidence BUT reasons are incomplete and/or part of the evidence is missing; OR only one complete & accurate perspective has been provided.
1	Response provides only subjective reasons (opinions) for choice and/or uses inaccurate or irrelevant evidence from the activity.	Response states at least one perspective of issue BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence.
0	No response; illegible response; response offers no reasons AND no evidence to support choice made.	No response; illegible response; response lacks reasons AND offers no evidence to support decision made.
X	Student had no opportunity to respond.	

Figure 2. Wilson et al. (2000) scoring guide

3. Methodology

3.1 Conceptual research framework

The research will be split into three main phases and are as follows. Figure 3 gives a diagrammatic representation of the phases.

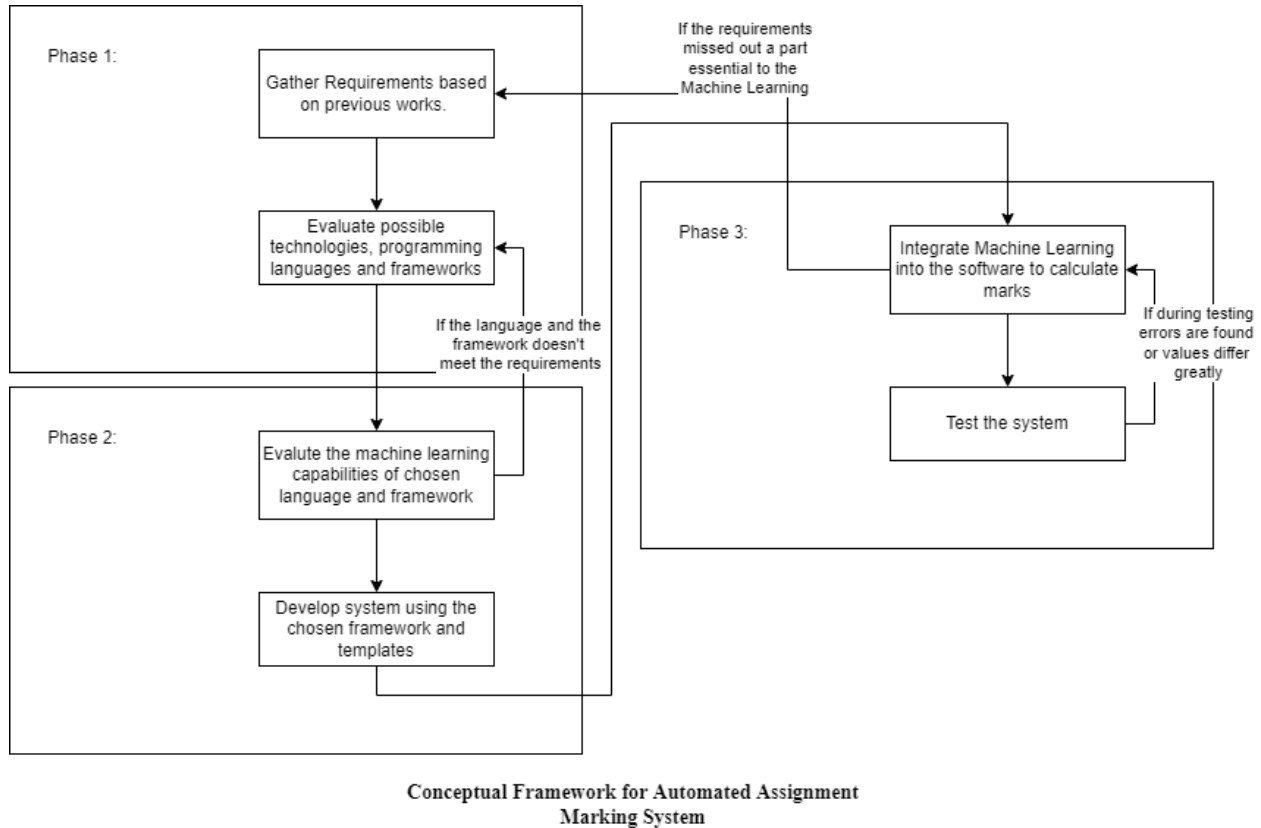


Figure 3. Conceptual Framework.

3.1.1 Phase 1

Phase 1 is the Requirement gathering phase and the framework choosing phase. During this phase I gathered the basic requirements of an assignment submission system without the evaluation of marks and then I added the Machine Learning to calculate marks automatically. The next aspect was to choose the programming language (Python, C++, C# or any other) to use and the platform the application would be deployed to as well as the specific technology to use to create the application (I had to pick between any of the following Winforms, React with .NET, WPF, Xamarin, Django, Android Studio with JAVA or with Kotlin).

3.1.2 Phase 2

Phase 2 is the evaluation of the chosen language and technology and whether or not it has the Machine Learning Libraries needed for the completion of the software. Owing to this I chose not to use any other language but Python due to the vast number of Libraries available and the versatility of it and the framework I chose was Django so I can create a web application which could be widely available. During this phase I also developed the system to be deployed using some templates and developing original parts to use to fulfil the requirements gathered in phase 1.

3.1.3 Phase 3

Phase 3 is the part where I added the Machine Learning code to the already complete system to compute the results. For this phase I used the Spacy ML Library which is mainly a library for Natural Language Processing and has already trained models and also is dependent on other Models such as NumPy and Doc2Vec and others to compute its result. After this aspect was done, I moved into testing of the system and any errors encountered were fixed the errors which stemmed directly from the requirement lacking a part were solved by modifying the requirements and then going through the phases again until the end phase has been reached.

3.2 Application methodology: Waterfall model

The methodology used for the creation of the application was the Waterfall model. This is a particular implementation of the software development life cycle that focuses on sequential development like a waterfall. This methodology has each phase completely wrapping up before the next phase begins. The waterfall model is highly dependent on a lot of work being done at each stage as there is no going back. It is efficient for small projects and can provide an effective release date.

3.3. Stages in the waterfall model

The waterfall model has 5 stages or phases and team members usually work independently on each stage though phases have to be completed in sequential order.

The stages namely Requirements, Design, Implementation, Verification and Testing are discussed below with the aid of Figure 4.



Figure 4. The waterfall model

3.3.1 Requirements

The project requirements have to be gathered and understood before any work can be commenced. The project requirements will be obtained from the stakeholders. This will be presented in the form of a document which contains details about each stage and also other important bits such as timelines, cost, risks and the success rates.

3.3.2 Design

The developers are required to design a technical solution based on the requirements. This is where things like scenarios, layouts and data models. This also where the scope of the project is identified.

3.3.3 Implementation

After the design has been completed this is where the technical implementation begins using hardware and software technologies. This is where the coding is done based on the requirements and specifications. Changes are usually minimal in this stage but if big changes need to be implemented then it's to go back to the design phase.

3.3.4 Verification or testing

Testing is done before the product or service can be released to the public, testing techniques such as white hat testing, black box testing and the like are done at this stage.

3.3.5 Deployment and maintenance

This is when the software is actually out for use and this is also when plans for future versions are made.

3.3.6 Advantages and disadvantages of the Waterfall Model

The waterfall modes are advantageous for many reasons one is because it helps system designers to find errors during the design and analysis stage and saves them the trouble in the implementation and testing stage. Another is that the cost and time for the software to be delivered can be estimated. Progress can be followed because the end of each stage is a milestone to reach. New developers can understand the project easily due to the extensive requirements document. Also, since it is not as iterative it is completed faster since the stakeholders aren't adding new unnecessary features.

The disadvantages though are that if the project is big, it will take a longer time to complete than the agile methodology or the V model. Clients don't usually know everything they want from a software at the start so they prefer to ask for changes during development and new features later down the road. This methodology also means the clients are not involved in the design and implementation stages. Finally, the biggest issue with this methodology is if one phase is delayed all the other phases are delayed.

3.3.7 Why do I use the Waterfall Model

I selected the waterfall model for the sole purpose that it is used for small programs that do not require large amounts of requirements. I also use it as this is a project that offers a clear intention of how the project will be done and how the software should look like from the get go. Also, the requirements of the program won't change as I go ahead and create the program. All the stages of the program would be outlined from inception to implementation. If this program does go on to be used by other institutions it would be better to use another model such as the V model.

3.4 Functional requirements

These are the things the system should do and the features the software should provide in order to gather these requirements. The stakeholders in this system are the users, the students and the lecturers.

Below are the functional requirements that have been obtained from the stakeholders mentioned above:

1. Allow students to login and logout.
2. Allow students to sign up.
3. Allow lecturers to login and logout.
4. Allow lecturers to sign up.
5. Allow lecturers to upload assignments.
6. Allow lecturers to upload sample answers.
7. Allow lecturers to view students' submissions.
8. Allow students to submit assignments for lecturers' course.
9. Allow lecturers to view grade of the assigned work.
10. Allow lecturers to delete assignments and submissions.
11. Allow lecturers to create and delete course.
12. Allow lecturers and students to edit details.
13. Allow users to email for newsletters and updates.

3.5 Non-functional requirements

These are the quality aspects of the system. They are not as essential to the functionality of the system but they are preferred.

1. Security

The system should be secure so whoever uses the system knows their details are not compromised. The system should be private and it should have integrity. It should be able to stop hackers from obtaining information about a customer from a system and then using it to exploit said customer for malicious or financial gain.

2. Usability

The system should be friendly and allow the users to interact with it comfortably. If a product has a lot of features and is not usable then the users will choose to go with another artifact or service that is easier to use and can be remembered intuitively.

3. Availability

The system should be available when the users need to use it and there shouldn't be downtime with any of the servers or the database. It should also be available on any of the web platforms.

4. Performance

The system should be responsive and fast and provide quick results. The system should not slow down unnecessarily or keep the users waiting and the system should be able to handle a high number of requests without downtime.

5. Accuracy

The systems should be able to provide accurate results for the user and the lecturer to see. The results should be consistent with what a teacher would award a student in every scenario and should provide consistency

6. Fault tolerance

The systems should be able to handle errors within the system and send the appropriate message to the users. The errors should be minimized from the developers and care should be taken so the system doesn't crash outright under unforeseen circumstances.

7. Efficiency

The system should be as efficient as possible and consume the least resources possible. The system should be able to use the least threads possible on the server to ensure that it runs optimally.

8. Cost

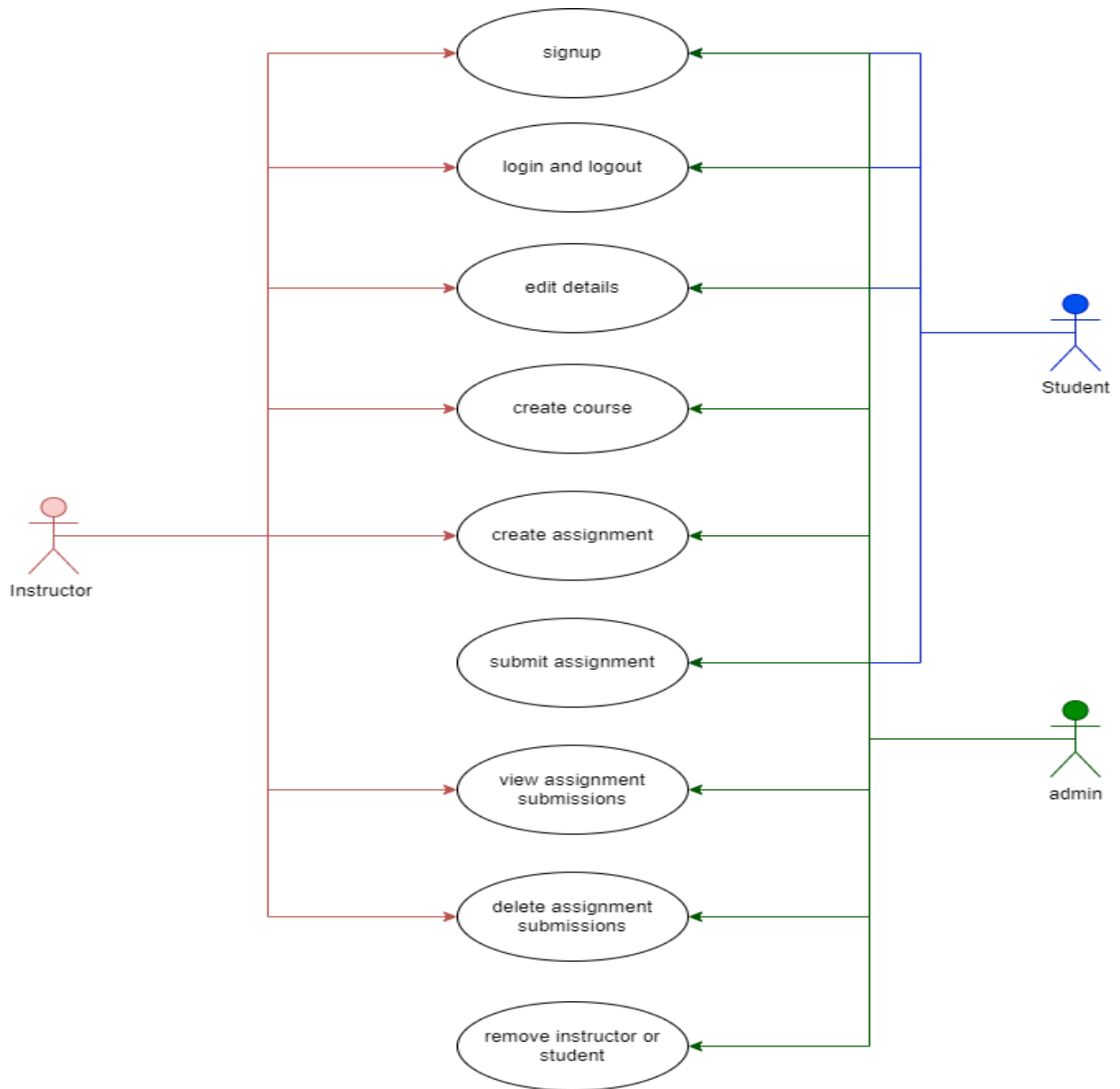
The system is a free to use systems so it is very cost effective from that standpoint. But generally, software's shouldn't be too expensive for the end user but should be appropriately priced so the users would be happy and the developers appropriately paid.

3.6 Tools used

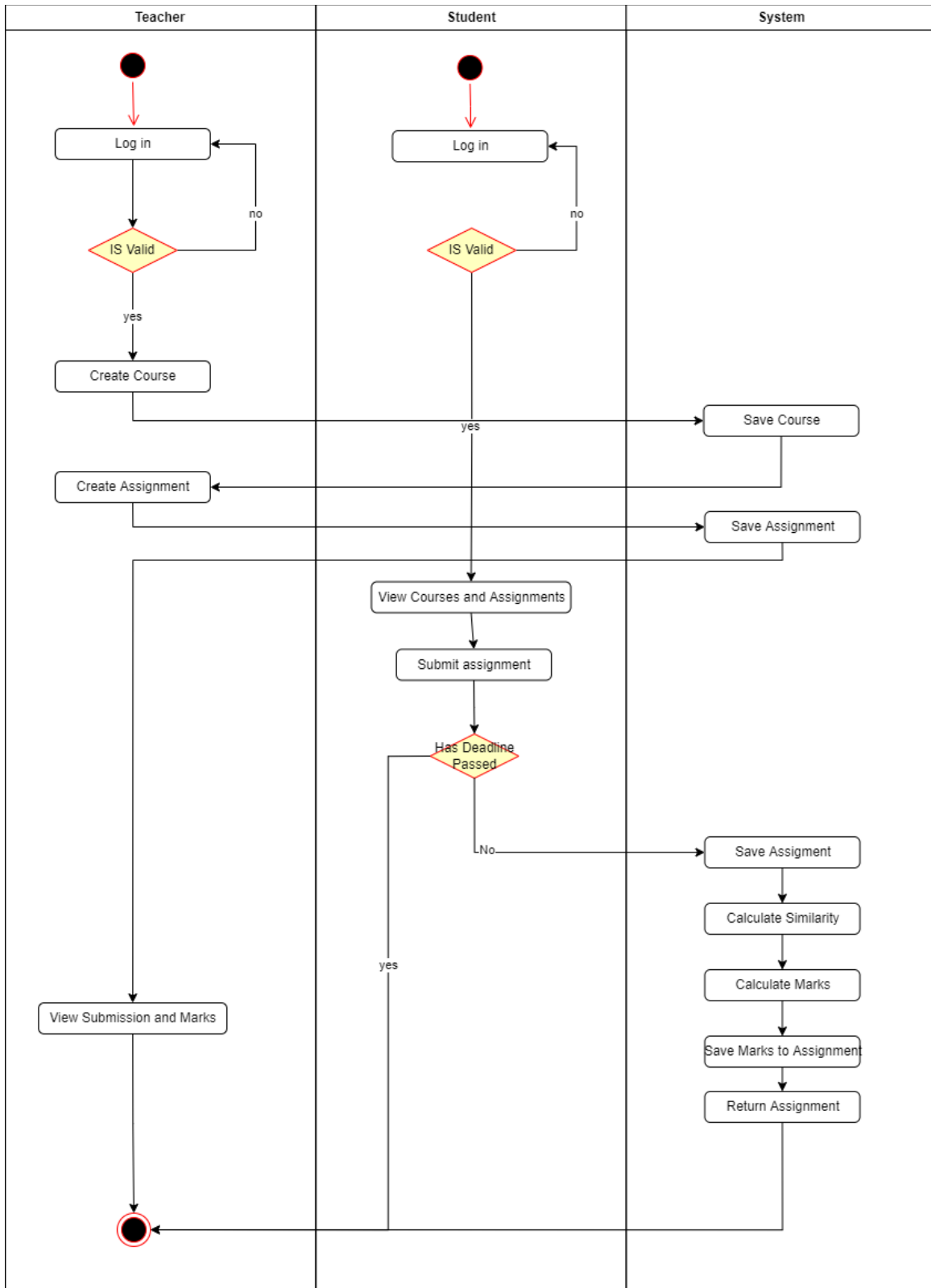
- 1) Visual Studio Code, <https://code.visualstudio.com> : A lightweight Integrated Development Engine developed by Microsoft for use with all languages and frameworks. It runs on windows, Mac OS and Linux as well.
- 2) Postman, <https://www.postman.com>: This is an API platform built for developers to create and test API's before they are used in development.
- 3) Browsers: Chrome and Edge to test the final outlook of the page.
- 4) Python, <https://www.python.org>: This is the main language that was used for the development of the application.
- 5) GitHub, <https://github.com/>: This is a code repository, where you can find templates and other projects by other developers as well as examples. It can also be used for collaboration between developers and project management.
- 6) Django, <https://www.djangoproject.com/>: This is the main framework used to develop this application, Django is based on python and supports dynamic pages as well as SQLite; this is a database written in languages and can be embedded in frameworks. It is present in android studio as well.
- 7) Stack Overflow, <https://stackoverflow.com/>: This is an open platform for developers to post their issues and get answers as well as pointers on what to do to solve their problems.
- 8) Quora, <https://www.quora.com/>: This is a question-and-answer social platforms designed for all sorts of questions.
- 9) Windows, <https://www.microsoft.com/en-gb/windows/?r=1>: This an Operating System, the most popular operating system in the world. It developed by Microsoft and it has multilingual support and was one of the first OS to come with a GUI (Graphical User Interface).
- 10) spaCy, <https://spacy.io/>: This is a very fast and easy to use software library for advanced natural language processing founded 7 years ago. It also be used in other frameworks such as TensorFlow and pyTorch.

3.7. Diagrams

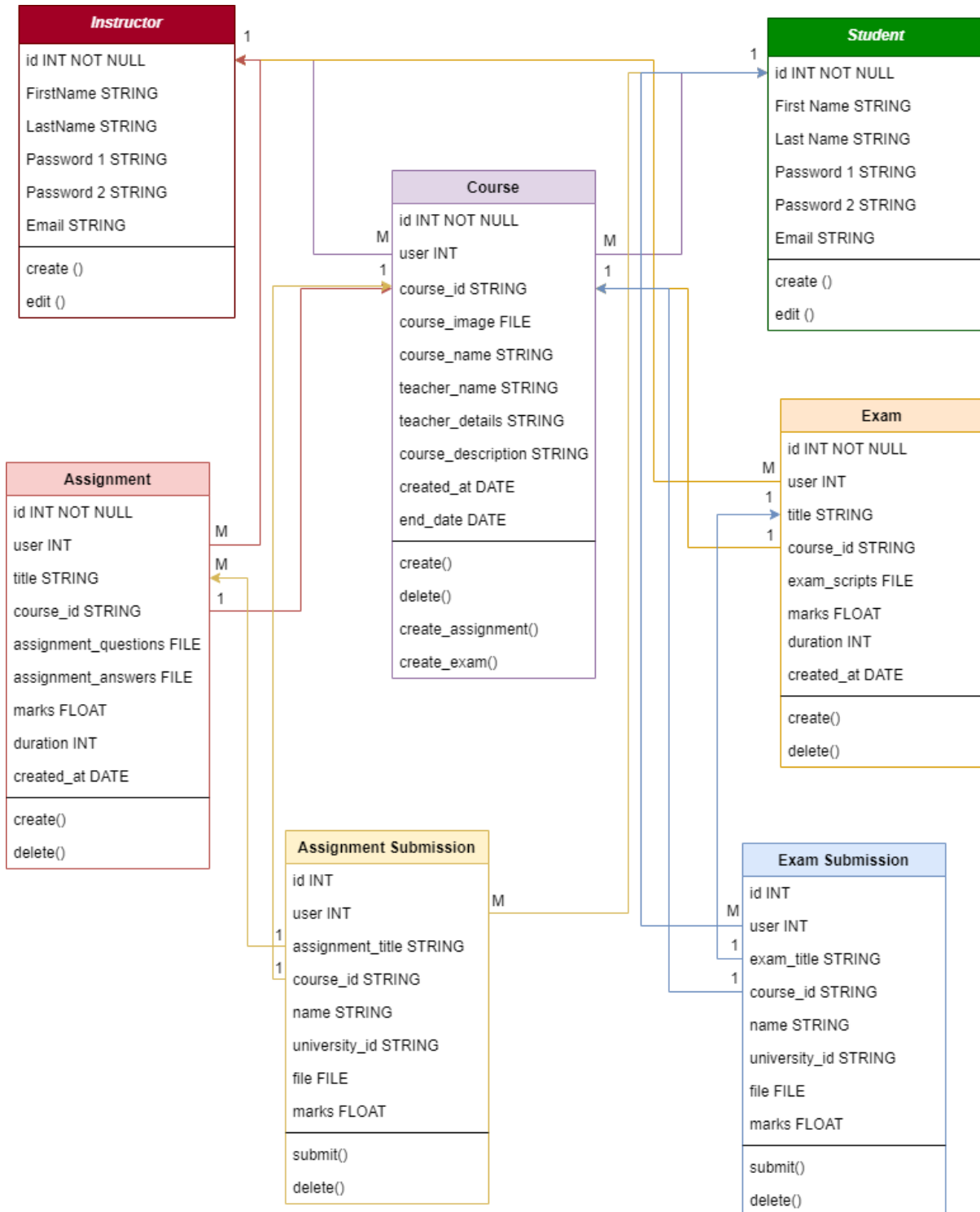
3.7.1 Use case diagram



3.7.2 Activity diagram

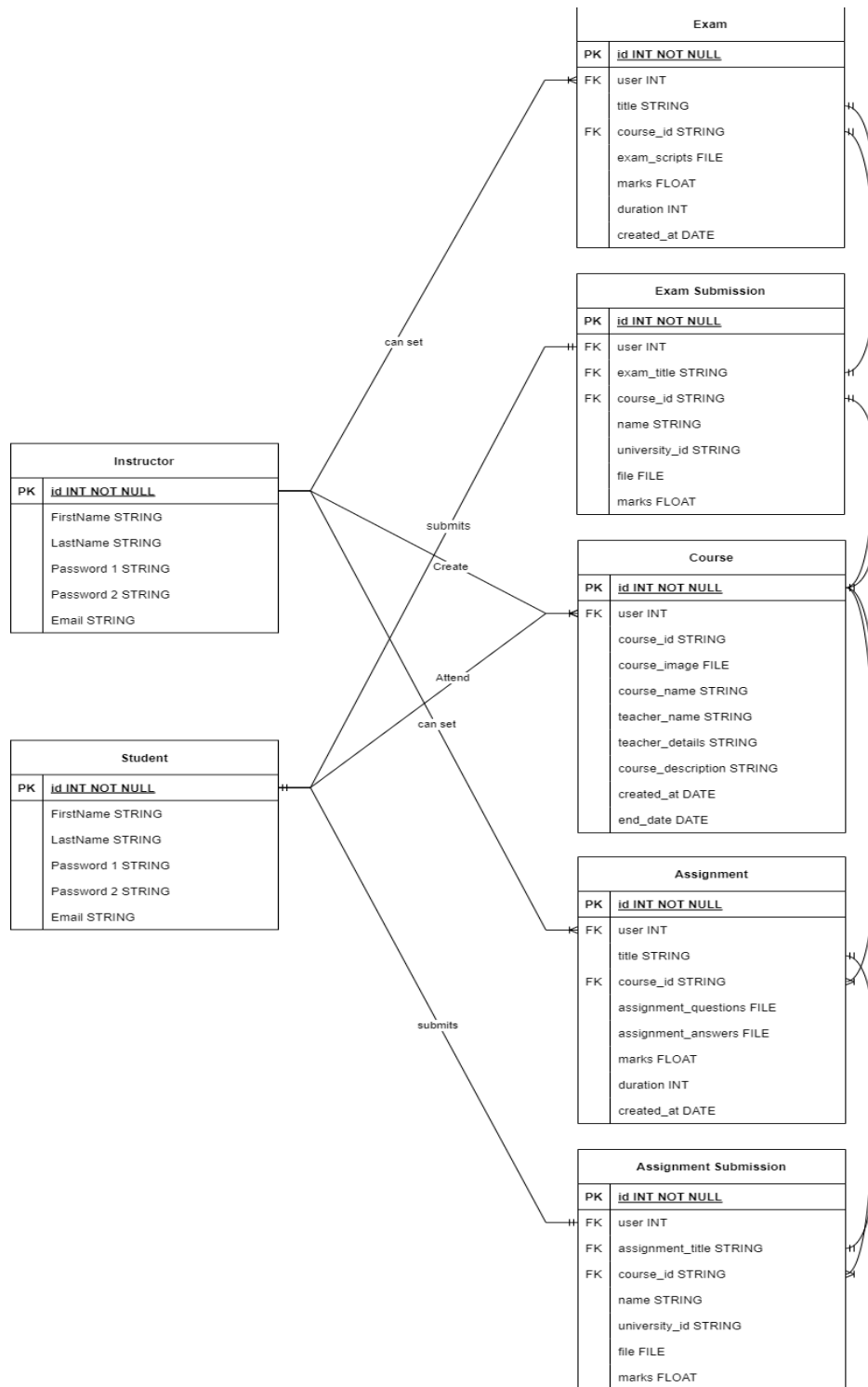


3.7.3 Class diagram



4. Implementation

4.1 Entity relation diagram



4.2 Testing

Testing was conducted with answers from Quora on the question: What is design thinking? The link to the page is placed below and the results of the program are specified below.

<https://www.quora.com/What-is-Design-Thinking>

The results of the program as well as the answers are put in Table 1 below.

Table 1.

	Answers	Score Awarded
Sample Answer	Design Thinking is following a human-centric approach while dealing with any problem. First identifying the problem which is affecting humans and then finding the solution. The very common issue with human behavior is that we always tend to directly jump on the solutions. We don't make efforts to find the real cause behind a problem (Context). Design Thinking helps us in finding the real cause keeping in mind three main aspects: People, Technology, and Business.	50
1.	There are various problems in this world. Many of them are complex and some of them are simple. We need to solve those problems. Since many solutions are available to a single problem but out of those many, we need to figure out the best one. The best solution is generally termed as the innovative and that is out of the box. Thinking divergently instead of convergently and to come up with a creative idea that might not exist for a problem that works well for the real problem, is all about Design thinking. In one line we can state "Design Thinking is a process for creative problem solving". Design thinking is a human-centered approach to innovation that draws from the designer's toolkit to integrate the needs of people, the possibilities of technology, and the requirements for business success."	28.786405880065214
2.	There are diverse issues on this world. Many of them are complicated and a number of them are simple. We want to remedy the ones issues. Since many answers are to be had to an unmarried hassle however out of these many, we want to determine out the pleasant one. The pleasant answer is commonly termed because the progressive and this is out of the box. Thinking divergently rather than convergently and to provide you with an innovative concept that may not exist for a hassle that works properly for the actual hassle, is all approximately Design wondering. In one line we are able to state "Design Thinking is a method for innovative hassle solving". Design wondering is a human-focused technique to innovation that attracts from the designer's toolkit to combine the desires of people, the opportunities of technology, and the necessities for enterprise success."	23.029742123351877

3.	<p>There are many problems in the world. Some of them are complex and some of them are simple. We need to find a way to solve those problems. Since many solutions are available for one problem but among those many, we need to find out the best one. The best solution is usually innovative and out of the box. Creative thinking is all about coming up with ideas that are different from the ones that are already out there. By thinking divergently, you can create new solutions to problems that work better than the ones that are currently being used. Design Thinking is a process for solving creative challenges. It is a human-centered approach that uses the design tools of a designer to integrate the needs of people, the possibilities of technology, and the requirements of business success.</p>	28.73406602683103
4.	<p>There are numerous issues on the planet. Some of them are mind boggling and some of them are straightforward. We really want to figure out how to take care of those issues. Since numerous arrangements are accessible for one issue yet among those many, we want to figure out the best one. The best arrangement is normally creative and out of the container. Innovative reasoning is tied in with concocting thoughts that are not the same as the ones that are now out there. By thinking differently, you can make new answers for issues that work better compared to the ones that are presently being utilized. Configuration Thinking is an interaction for settling imaginative difficulties. A human-focused approach utilizes the plan instruments of a creator to coordinate the necessities of individuals, the conceivable outcomes of innovation, and the prerequisites of business achievement.</p>	26.526140094197046
5.	<p>Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged.</p>	4.3467948056326335

5. Results and findings

5.1 Summary and conclusion

After the development of the program and testing we see that the data in numbers 1 through to 4 entered even though they are small paragraphs more or less mean the same thing as the sample answer but the spaCy model returned the scores to range from 23-28 out of 50 marks. This means the answers were half correct according to the models. These same answers which when compared by a human are more or less the same thing. The only exception here is the 5th answer with 'Lorem Ipsum' this was thrown to intentionally test the system and the system responded appropriately and awarded it 4.34 out of 50 marks. This means the above answers that got more than half marks are considered somewhat similar according to the system. This correlates with Luckin (2017) who stated estimated that "an annual budget of US\$600 million per year" would be needed for a complex AI project. The spaCy model is free and open Source so there could be issues whether the NLP Libraries are really as advanced as they should be. In its defense though the similarity function was created to handle single sentences instead of large paragraphs.

5.2 Recommendations

My first recommendation to improve the validity of this software is to use the largest model offered by spaCy, due to space constraints I used the medium model which was 400MB for the libraries alone and a further 1.4GB for everything else added on.

The next recommendation is to use a dedicated paid API designed to handle paragraphs so the development of the system can be quick.

My third and final recommendation is that proper research over a period of time should be put into creating an AI model specifically for marking essays and one that can do it by itself without the aid of anything else such as sample answers or marking scripts.

5.3 Limitations

This research will be limited by the fact that the system can only handle the marking of short sentences accurately and not long paragraphs. The system is also limited by the fact that it can only mark with the aid of the marking scheme and not without it so it is not a truly intelligent model in that regard. The system will be limited by the fact that it can only compare answers in English and not in other languages or disciplines. Another limitation is the fact that the documents have to only contain the answers and no other additions such as a cover page for the system to work effectively.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

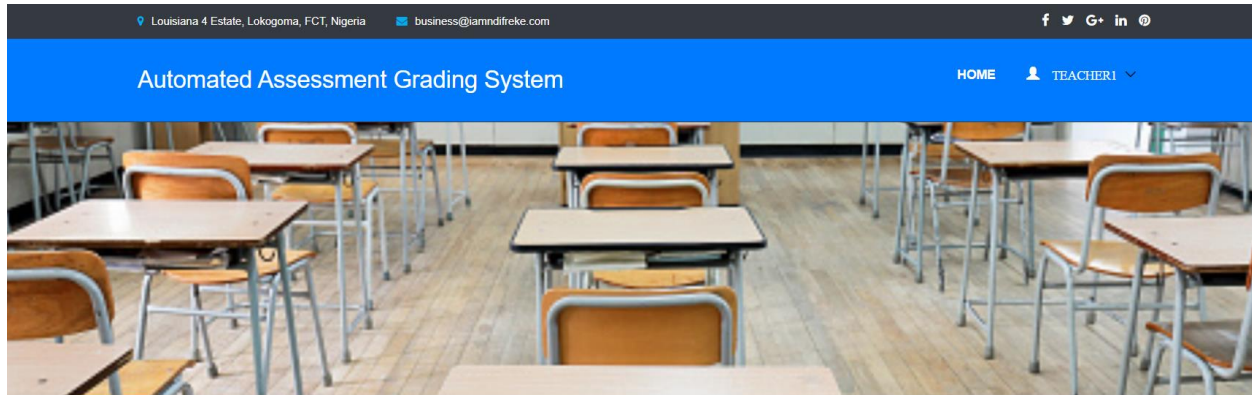
References

- Adams, W. L. (1932). Why teachers say they fail pupils. *Educational Administration and Supervision*, XVIII, 594-600.
- Ana, P., & Tawo Bukie, P. (2013). *Design and implementation of online examination administration system for universities*. <https://www.researchgate.net/publication/327075106>.
- Archer, A., & B. McCarthy (1988). *Personal biases in student assessment*. <https://doi.org/10.1080/0013188880300208>
- Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.
- Bloxham, S. (2009). *Marking and moderation in the UK: false assumptions and wasted resources*. <https://doi.org/10.1080/02602930801955978>
- Boussakuk, M., Bouchboua, A., El Ghazi, M., & El Bekkali, M. (2021). *Designing and developing e-assessment delivery system under IMS QTI ver.2.2 specification*.
- Dikli, S. (2006). Automated essay scoring. *Turkish Online Journal of Distance Education-TOJDE*, 7(1), Article: 5.
- Elliot, S. (2000a). *A study of expert scoring and IntelliMetric scoring accuracy for dimensional scoring of grade 11 student writing responses (RB-397)*. Newtown, PA: Vantage Learning.

- Elliot, S. (2000b). *A true score study of IntelliMetric accuracy for holistic and dimensional scoring of college entry-level writing program* (RB-407). Newtown, PA: Vantage Learning.
- Elliot, S. (2001a). *About IntelliMetric* (PB-540). Newtown, PA: Vantage Learning.
- Elliot, S. (2001c). *Applying IntelliMetric Technology to the scoring of 3rd and 8th grade standardized writing assessments* (RB-524). Newtown, PA: Vantage Learning, 61.
- Elliot, S. (2002). *A study of expert scoring, standard human scoring and IntelliMetric scoring accuracy for statewide eighth grade writing responses* (RB-726). Newtown, PA: Vantage Learning.
- Elliot, S. (2003a). *A true score study of 11th grade student writing responses using IntelliMetric Version 9.0* (RB-786). Newtown, PA: Vantage Learning.
- Elliot, S. (2003b). *Assessing the accuracy of IntelliMetric for scoring a district-wide writing assessment* (RB-806). Newtown, PA: Vantage Learning.
- Elliot, S. (2003c). *How does IntelliMetric score essay responses?* (RB-929). Newtown, PA: Vantage Learning.
- Elliot, S. (2003d). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.). *Automated essay scoring: A cross disciplinary approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fisher, M. R. Jr. (2022). *Student assessment in teaching and learning*. <https://cft.vanderbilt.edu/student-assessment-in-teaching-and-learning/>.
- Hill, P., & Barber, M. (2014). *Preparing for a renaissance in assessment*. London: Pearson Education.
- James, D. A., & Gabunada Lambating, J. (2001). *Validity and Reliability in Assessment and Grading: Perspectives of preservice and in-service teachers and teacher education professors*. Conference of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
- Kuisma, R. (1999). *Criteria referenced marking of written assignments*. <https://doi.org/10.1080/0260293990240103>
- Kukich, K. (2006). Beyond automated essay scoring. *The Turkish Online Journal of Distance Education*. <https://www.semanticscholar.org/paper/Beyond-Automated-Essay-Scoring-Kukich/66a03e431c858ed3dd00d25773e3a2c9b5528e6c>.
- Luckin, R. (2017). *Towards artificial intelligence-based assessment systems*. <http://dx.doi.org/10.1038/s41562-016-0028>
- Luckin, R., & Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education, London.
- McKinstry, B., Cameron, H., Elton, R., & Riley, S. (2004). Leniency and halo effects in marking undergraduate short research projects. *BMC Medical Education*, 4, 28-28.
- Nottingham, M. (1988). *Grading practices – Watching out for land mines*. NASSP.
- Schinske, J., & Tanner, K. (2014). Teaching more by grading less (or differently). *CBE—Life Sciences Education*, 13, 159-166.
- Tomkinson, B., & Freeman, J. (2011). *Problems of assessment*. Conference: *International Conference on Engineering Education* (ICEE-2011). Belfast, Northern Ireland. https://www.researchgate.net/publication/273455702_Problems_of_Assessment.
- Willey, K., & Gardner, A. (2010). *Improving the standard and consistency of multi-tutor grading in large classes*. <https://www.uts.edu.au/sites/default/files/Willey.pdf>.
- Wilson, M., & Sloane, K. (2010). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13, 118-201. https://doi.org/10.1207/S15324818AME1302_4

Appendix

Home Page



Available Courses



Available Courses

Web Development | [VIEW COURSE](#)
Mr Mister

Design Thinking | [VIEW COURSE](#)
Mr Example


View Course Page

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@amndifreko.com | f t G+ in

Automated Assessment Grading System | HOME | TEACHER |

Web Development

CSC311



Instructor : Mr Mister

About Instructor
Testing 2

About Course
Testing 2

Assignments

Exams

Course Duration
July 14, 2022 - 2022-07-14

Automated Assignment Grading System

Marking should be automated and fair

Inquiries
For more inquiries about this system!

Email Address

View Assignment

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@amndifreko.com | f t G+ in

Automated Assessment Grading System | HOME | TEACHER |

Sample Assignment

Content

Marks
50.0

Time
2 hours

Delete Assignment

Create a Assignment

Automated Assignment Grading System

Marking should be automated and fair

Inquiries
For more inquiries about this system!

Email Address

Create Assignment Page

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@jamndifreke.com | f t G+ in

Automated Assessment Grading System | HOME | TEACHER1

Create Assignment

Assignment Name

Course ID

AssignmentQuestions
 No file chosen

AssignmentAnswers
 No file chosen

Marks

Duration

Automated Assignment Grading System

Marking should be automated and fair
Louisiana 4 Estate, Lokogoma, FCT, Nigeria

Inquiries
For more inquiries about this system!

Email Address

Edit Profile Page

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@jamndifreke.com | f t G+ in

Automated Assessment Grading System | HOME | TEACHER1

Edit Profile

First name

Last name

Email

Automated Assignment Grading System

Marking should be automated and fair

Inquiries
For more inquiries about this system!

View Assignments Page

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@jamndifreke.com | f t G+ In @

Automated Assessment Grading System | HOME | TEACHER1

Assignments

Name Ndifreke James Okpo
Id VUG/CSC/19/1000
Marks 28.788406880085214
Delete Submission
Name Ndifreke James Okpo
Id VUG/CSC/19/2000
Marks 23.029742123351877
Delete Submission
Name Ndifreke James Okpo
Id VUG/CSC/19/3000
Marks 28.73406602883103
Delete Submission

Delete Assignments Page

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@jamndifreke.com | f t G+ In @

Automated Assessment Grading System | HOME | TEACHER1

Are you sure you want to delete this Assignment?
Sample Assignment by VUG/CSC/19/1000
for Web Development

[Delete](#) [Cancel](#)

Automated Assignment Grading System | Inquiries
For more inquiries about this system!

Submit Assignments Page

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@iamndifreke.com | f t G+ in @

Automated Assessment Grading System | HOME COURSES STUDENT1

Submit Assignment

Course ID

AssignmentTitle

Name

University Id

Upload File
 No file chosen

Signup Page

Louisiana 4 Estate, Lokogoma, FCT, Nigeria | business@iamndifreke.com | f t G+ in @

Automated Assessment Grading System | HOME REGISTER LOGIN

Instructor Registration

First Name

Last Name

Email

Password

Confirm Password

Django Admin

Django administration

Site administration

AUTHENTICATION AND AUTHORIZATION	
Groups	+ Add ✎ Change

CORE	
Assignment submissions	+ Add ✎ Change
Assignments	+ Add ✎ Change
Courses	+ Add ✎ Change
Exam submissions	+ Add ✎ Change
Exams	+ Add ✎ Change

Recent actions

My actions

- ✘ VUG/CSC/19/1000**
Assignment submission
- ✘ VUG/CSC/19/2000**
Assignment submission
- ✘ VUG/CSC/19/3000**
Assignment submission
- ✘ VUG/CSC/19/4000**
Assignment submission
- ✘ VUG/CSC/19/5000**
Assignment submission
- ✘ VUG/CSC/19/1000**
Assignment submission
- ✘ VUG/CSC/19/4000**
Assignment submission
- ✘ VUG/CSC/19/3000**
Assignment submission
- ✘ VUG/CSC/19/2000**
Assignment submission
- ✘ VUG/CSC/19/5000**
Assignment submission



Detecting Phishing Emails Using Random Forest and AdaBoost Classifier Model

Fredrick Nthurima, Abraham Mutua & Waithaka Stephen Titus

Kenyatta University, School of Pure and Applied Science, Nairobi, KENYA

Received: 7 July 2023 ▪ Revised: 4 October 2023 ▪ Accepted: 15 November 2023

Abstract

Phishing attack occurs when a phishing email which is a legitimate-looking email, designed to lure the recipient into believing that it is a genuine email to open and click malicious links embedded into the email. This leads to user reveal sensitive information such as credit card number, usernames or passwords to the attacker thereby gaining entry into the compromised account. Online surveys have put phishing attack as the leading attack for web content mostly targeting financial institutions. According to a survey conducted by Ponemon Institute LLC 2017, the loss due to phishing attack is about \$1.5 billion per year. This is a global threat to information security and it's on the rise due to IoT (Internet of Things) and thus requires a better phishing detection mechanism to mitigate these loses and reputation injury. This research paper explores and reports the use of a combination of machine learning algorithms; Random Forest and AdaBoost and use of more phishing email features in improving the accuracy of phishing detection and prevention. This project will explore the existing phishing methods, investigate the effect of combining two machine learning algorithms to detect and prevent phishing attacks, design and develop a supervised classifier which can detect phishing and prevent phishing emails and test the model with existing data. A dataset consisting of both benign and phishing emails will be used to conduct a supervised learning by the model. Expected accuracy is 99.9%, False Negative (FN) and False Positive (FP) rates of 0.1% and below.

Keywords: classification, algorithm, cyber security, machine learning, spam emails, cyber security, cyberattack, web attacks, intrusion detection and phishing emails, AdaBoost, Random Forest.

1. Introduction

1.1 *Background to the study*

According to Anti-Phishing Working Group report 2018, Phishing attack is the number one attack committed by threat actors as compared to other attacks. It is a form of fraud where the attacker deceives the target for personal gain or reputation damage. Fraud results to users revealing their personal details like credit card numbers, passwords, PIN, usernames and other sensitive information leading to compromise of account and loss of funds.

Phishing campaign lures users to giving confidential information by visiting websites that have been made to look like legitimate websites (phishing.org, 2018). Phishing is carried out using a digital garget like a computer or iPad through a computer network. Malicious actors usually target weakest element in the security chain, i.e., end-users (Khonji, Jones & Iraqi, 2013).

Attackers in the phishing campaign usually craft messages known as social engineered messages persuading users to click and visit the illegitimate websites thus revealing their confidential information to attackers. This enables threat actors gain entry into the compromised account and achieve their objectives like data theft, funds transfer or reputation injury.

For instance, a malicious email might have a malware which when clicked by the user will install itself in the pc or mobile phone and will transfer funds to the account of the attacker whenever the owner of the account tries to transfer cash (Khonji et al., 2013). This attack is called Man in the browser (MITB) which is a variant of the Man in the middle (MITM) attack. The man in the browser attack usually uses different vectors like ActiveX components, plugins or email attachments to deliver the payload to the user's computer or phone.

With the increasing case of cyber-attacks, organizations are looking for safer ways of protecting data and prevent getting hacking or getting hacked again. Design and technology should be greatly improved to ensure hackers do not infiltrate into networks.

According to (Behdad, French, Bennamoun & Barone, 2012), using better defense systems is not enough in stopping malicious actors from penetrating systems since these are sometimes circumvented; a better system should detect malicious activities and prevent them before causing any damage.

There are a number of mechanisms used today to filter spams but they are static in nature such that they cannot handle the ever-evolving threats and phishing trends. They are only capable of detecting already known phishing patterns leaving behind future attacks. This is a security weakness because attackers are not static in nature and use different ways of evading detection. This challenge has motivated researchers into looking for other ways of detecting both known and new threats which led to the knowledge and use of machine learning algorithms.

Machine learning (ML) is a discipline of artificial intelligence that uses data mining to detect new and existing phishing features from a given dataset which is ultimately used for classification of benign and phishing emails.

In this project, we will use a combination of two ML algorithms namely random forest and AdaBoost and a set of 15 important phishing features as identified from the literature. The dataset will consist of 3000 emails from both phishing and benign sources and then extract the features for each email, form a vector representation of these extracted features which will be used to train our classifier model.

1.2 Problem statement

Very few phishing email filters have been developed as opposed to many existing email filters that have been developed for spam emails. Many of them used several phishing detection techniques ranging from blacklists, visual similarity, heuristic, and machine learning. Of all these techniques, ML-based technique does offer the best results (Brown, Ofoghi, Ma & Watters, 2017).

However, current machine learning anti-phishing solutions use a single algorithm to detect phishing. This according to results doesn't offer best accuracy of detection which currently stands at 98% (Smadi, Aslam, Zhang, Alasem & Hossain, 2015). Moreover, they have used domain/url characteristics leaving behind other phishing features that are present in phishing emails therefore lowering accuracy and detection rates.

There is need to investigate the use of combination of two machine learning algorithms namely Random Forest and AdaBoost and include other phishing email features to increase detection accuracy to 99.9%, fewer FPs and FNs and increase overall phishing detection and prevention.

1.3 Objectives

This project is aimed at achieving the following objectives:

1. To investigate the existing phishing attacks methods used by attackers to lure users.
2. To investigate the effect of combining Random Forest and AdaBoost algorithms and use of more features in phishing email detection.
3. To design and develop a supervised classifier model which can detect phishing emails.
4. To test the classifier model with existing data.

1.4 Research questions

1. How do attackers lure users to visit phishing websites?
2. Can the use of combination of ML algorithms and use of more features lead to increased accuracy in phishing detection?
3. To what accuracy can ML achieve phishing detection?
4. What recommendations can be inferred for future classifiers?

1.5 Research scope

1. This project will address phishing emails.
2. The project will use a combination of Random Forest and AdaBoost machine learning algorithms.
3. A total of 15 learning features will be used.
4. Algorithms will be implemented using python frameworks.

2. Literature review

2.1 Introduction

Phishing attack occurs when a phishing email which is a legitimate-looking email which is designed to lure the recipient into believing that it is a genuine email to open, and click malicious links embedded into the email. This leads to user revealing sensitive information such as credit card number, usernames or passwords to the attacker thereby gaining entry into the compromised account (Holbrook, Kumaraguru, Downs, Cranor & Sheng, 2010).

Approximately 57% of phishing attacks target financial institutions and payment services, according to Phishing Activity Trends Report – 4th Quarter 2017, Anti-Phishing Working Group (APWG).

Phishing is a widely spread threat in the Internet and is achieved when an attacker lures a user into entering sensitive information like passwords or credit card numbers into illegitimate website that is controlled by malicious actor. It has been demonstrated that social phishing, where the word “social” means information related to the victim is used, produces very effective results compared to regular phishing. Gupta, Prakash, Kompella and Kumar (2015) found that if phishing e-mails impersonated a target’s friend, the success rate of the phishing attack

increased from 16% to 72%. The social aspect of information is therefore not only of value to social network operator but also to attackers. This is made even more possible if the information on social media contains a valid email address or there is a recent conversation between the victim and the impersonated friend.

With automation of data extraction from social media networks, a lot of usable data is available to attackers which can be used to carry out phishing attacks. Information extracted from social media networks is misused by the context-aware spam to increase appearance of authenticity of traditional spam messages.

Brown, Ofoghi, Ma and Watters (2017) classified three context-aware spam attacks: relationship-based attacks, unshared-attribute attacks, and shared-attribute attacks. Relationship-based attacks exploit relationship information only thus making this the spam equivalent of social phishing. The other two attacks exploit additional information from social networks, information that is either shared or not shared between the spam target and the spoofed friend. An example of an unshared attack are birthday cards that seem to originate from the target's friend. Shared attributes, e.g., photos in which both the spam target and the impersonated friend are tagged, can be exploited for context-aware spam. Huber, Mulazzani, Leithner, Schrittwieser, Wondracek and Weippl (2011) found that the missing support for communication security can be exploited to automatically extract personal information from online social networks. Furthermore, the authors showed that the extracted information could be misused to target a large number of users with context-aware spam.

Gupta, Prakash, Kompella and Kumar (2015) used a hybrid of two techniques namely blacklists and heuristics to detect phishing emails which achieved a FP and FN rates of 5% and 3% respectively.

Holbrook et al. (2010) conducted an investigation on some anti-phishing toolbar and reported SpoofGuard which was developed by Ledesma, Chou, Mitchell and Teraguchi (2014) to have a FP rate of 38% and a FN rate of 9%. Also, Nargundkar, Tiruthani and Yu (2017) developed a heuristics-based phishing detection system which achieved a FP and FN rates of 1% and 20%, respectively. Smadi et al. (2015) also used heuristics technique and their method achieved a FP rate of 3% and FN rate of 11%.

Sadeh, Fette and Tomasic (2017) used Machine Learning based technique and they achieved a FP rate of 1% and a FN rate of 1.2%. Strobel, Glahn, Moens, De Beer and Bergholz (2010) combined the use of heuristics and ML-technique and their method achieved a FP rate of 0.05% and a FN rate of 1%.

All these proposed methods have relatively high FP rate and FN rate except for Sadeh et al. (2017) and Strobel et al. (2010) whose techniques achieved excellent results with very low FP and FN rates. However, Strobel et al. (2010) used model-based features involving the processing of images which results in increased runtime and space. Sadeh et al. (2017) also made use of a domain name feature that has to be obtained by sending of queries over the network which results to increased run-time. In our proposed method, the phishing email features are extracted directly from the email. Thus, by eliminating sending of queries, the proposed model will be faster and remove space complexities.

2.2 Common ml anti-phishing techniques

Phishing emails can be classified using complex techniques based on specific features such as URL length, sub_domain, prefix_suffix and many more. Mohammad, Thabtah and McCluskey (2013) created unique learning bases making use of space understanding to detect phishing and legitimate emails. Recently, there has been many studies for achieving automated

rules to separate genuine and phishing emails with the use of statistical analysis (Abdelhamid, Thabtah & Ayes, 2014). For instance, Mohammad, Thabtah and McCluskey (2014) grouped many intelligently derived rules in regards to different phishing features by using frequency counting of phishing emails (instances) gathered from various sources including PhishTank and Yahoo directory. Improvements in rules for decision making have been developed whereby a computational intelligence method on a larger phishing dataset collected from various sources have been used (Abdelhamid, Thabtah & Ayes, 2014).

Phishing was studied using C4.5, decision tree, random forest, support vector machine and Naïve Bayes approaches. “Phishing Identification by Learning on Features of Email Received” (PILFER) was developed as an anti-phishing technique and then investigated on a set of 860 phishing case and 695 ham cases. The results were different features for recognising instances as phishing or ham, i.e., IP URLs, time of space, HTML messages, number of associations inside the email, JavaScript and others. Therefore, the authors explained that PILFER can improve the clustering of messages by joining all ten features found in the classifier beside “Spam filter output”.

In order to reduce both false positives and false negatives an evaluation of Random Forest algorithm was conducted against 2000 messages (Akinyelu & Adewumi, 2014). After experimentations with a 15-feature dataset, the results show a reduction in error rate when using Random Forest and therefore use of this algorithm as a method for phishing classification seemed fitting. The models using Random Forest seemed to be more dominant with respect to detection rate.

Aburrou, Hossain, Dahal and Thabtah (2010) conducted another project to accurately classify websites based on features. The authors manually classified features into six criteria and then load them into an environment for analysis on Waikato Environment for Knowledge Analysis (WEKA). During this exercise, various experiments were ran using four classification algorithms against 1006 instances from PhishTank. The evaluation criteria to determine the applicability of the features was the classification accuracy. The results showed that decision tree algorithms achieved detection rate of an average of 83% of the phishing sites. The authors proposed that with use of appropriate pre-processing, the detection accuracy would improve.

Enhanced Dynamic Rule Induction (eDRI), is one of the first Covering algorithms that has been applied as an anti-phishing tool (Thabtah, Qabajeh & Chiclana, 2016). This Covering algorithm processes datasets by using two main thresholds, frequency and Rule strength. eDRI scans the training dataset and only stores “strong” features if their frequency exceeds the minimum frequency threshold. As a result, all these features become part of the rule while all other values are removed during the initial scan. Once a rule is derived, eDRI removes its training instances and updates the strong features frequency to reflect the removal of its instances. Hence, eDRI somehow naturally prunes features and leads to a more controllable models. As part of the experiments, 11,000 websites were collected from multiple sources to evaluate eDRI’s reliability. In comparison decision tree algorithm, the results acquired showed eDRI superiority to other Covering and decision tree approaches with respect to phishing detection rate.

A machine learning technique that has been highly criticised as a result of its time consumption in tuning its parameters is trial and error Neural Networks (Mohammad, Thabtah & McCluskey, 2013). This technique usually requires a domain expert available during the parameter tuning stage. A Neural Network anti-phishing model proposed the elimination of trial and error and aimed for a more self-structuring classification (Thabtah, Mohammad & McCluskey, 2016). The authors designed the self-structured approach by updating several parameters, like the learning rate dynamically before adding a new neuron to the hidden layer. Therefore, the process of updating the NN features is performed while building the classifier in the network environment. The purpose of applying the dynamic NN model was to detect phishing instances from a real

dataset found in the UCI data repository using different epoch sizes (100, 200, 500, 1000) (Mohammad, Thabtah & McCluskey, 2015). The results revealed promising predictions when compared to Bayesian networks and decision trees.

Phishers keep on updating their deceptive methods hence there was need to develop an anti-phishing NN model that is based on constantly improving the learnt predictive model based on previous training experiences (Mohammad, Thabtah & McCluskey, 2014). The goal was to cope with the aggressive efforts by phishers that frequently update deceptive methods, therefore developed a self-structuring NN classification algorithm that deals with vitality of phishing features. This self-structuring NN algorithm uses validation data to keep track on the performance of the built network model and involves appropriate intelligent decisions based on the outcomes acquired against the validation set. For instance, when the attained error against the network is less than the minimum achieved error so far, the algorithm saves the networks' weights and continues the training process. On the other hand, when the achieved error is larger than the minimum achieved error so far, the algorithm continues the training process without saving the weights. Moreover, if need be, updates on other important network parameters occur during the construction of the classifier without having to wait until the model has been entirely built. As part of the experimentation on a number of features dataset revealed that the self-structuring NN model was able to generate highly predictive anti-phishing models compared to traditional classification approaches, such as C4.5 and probabilistic approaches.

2.3 Training and testing the model

The Validation and Testing modules of the NN model includes two components of “Sample Matrix” and “Output Matrix” as follows.

- “Sample Matrix”: this matrix contains sample data from the “Input Matrix”. The trained NN model uses the data in the “Sample Matrix” as inputs during the testing phase. In our implementation, this matrix is a logical $n \times 5$ matrix contains n sample data from the

“Input Matrix”.

- “Output Matrix”: this matrix contains output data for the data in the “Sample Matrix”. The trained NN model predicts the output values for the “Sample Matrix” and stores them in the “Output Matrix”. In our implementation, this matrix is a logical $n \times 1$ matrix contains output data for the emails represented in “Sample Matrix”. The trained NN model predicts the output value, in terms of an email being benign or phishing, for each email in the “Sample Matrix”. These predictions will be stored in the “Output Matrix” and will be used to evaluate the performance of the neural network. We used 70% of the entire dataset, which includes all the benign emails from PhishTank dataset and all the phishing emails from PhishCorpus dataset, for training, 15% for validation and 15% for testing. We used scikit learn framework to develop, train, validate and then test our classifier. model. Our developed NN model has 10 hidden layers, 5 input features, 1 output layer, and 1 output features, Figure 3. The captured results are discussed in the next section.

Scikit-learn *KFold* class will be used to automatically implement k-fold cross-validation on the given data set.

3. Methodology

3.1 Introduction

Machine learning is made up of training phase and the testing phases. We intent to use two datasets to train our model; benign and phishing emails.

We will obtain the datasets sets from Alexa for Benign emails and PhishTank for phishing emails. We intend to use a combination of two algorithms;

1. Random Forest;
2. AdaBoost; to increase accuracy of RF.

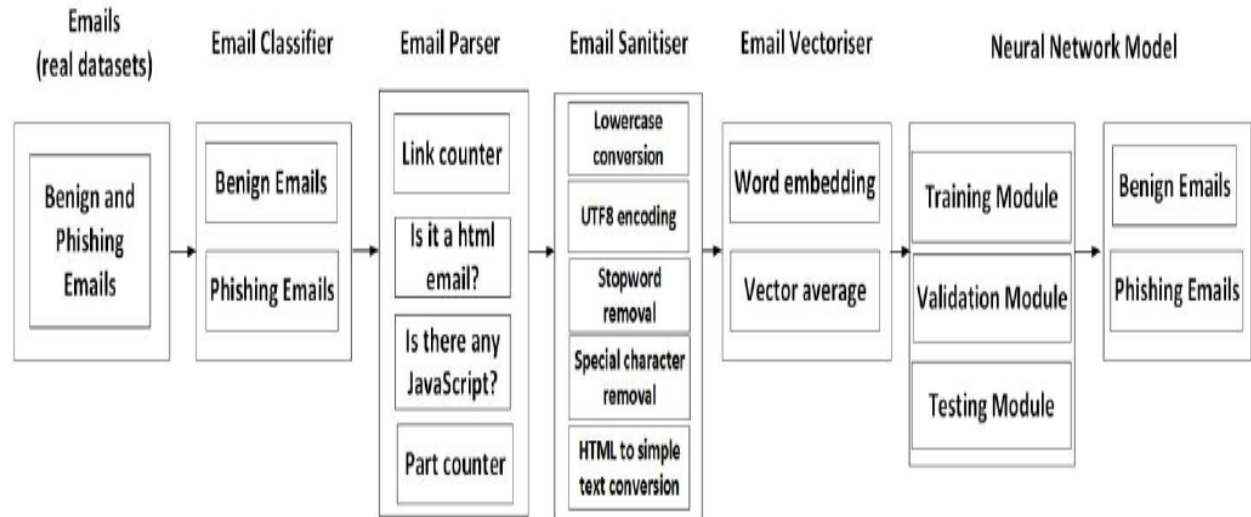


Figure 1. ML classifier modelling of the proposal

Tools to assist in data modeling

- Open source threat data training set called EMBER.
- IBM Watson
- Existing Anti-Phishing solutions like Spam Assassin.
- PhishTank and Alexa for data sets

Programming languages

- Python using libraries like *scikit-learn*, *pandas*, *numpy* and *matplotlib*
- Java

To train and test our classier, we will use a method called 10-fold cross validation. In this method, the training dataset is divided into 10 parts; 9 of the 10 parts will be used to train our classifier and the information obtained from the training phase will be used to validate (test) the 10th part. This process is repeated 10 times in such a way that at the end of the training and testing processes each of the parts will be used as both training and testing data. The cross validation technique ensures that training data id different from test data. In the area of machine learning, this method has shown to provide very good estimate of error of classifier.

3.2 Model features

Features used in the email classification

This section describes the phishing features that our classifier will use. These features are identified from different literature thus forming a combination set of features that effectively classify phishing and non-phishing emails.

This project will to use a total of 15 features identified from different literature commonly used by phishing attackers. These features are described below.

IP-based URLs

Legitimate websites normally contain the name of the website on the URL for example <http://www.forexample.com/>, tells the user one wants to connect to the website of forexample. Attackers usually mask their identity by replacing the domain name with IP address, e.g., <http://42.56.100.21/login.asp>. By doing so the attackers are able to evade detection by using IP-based URLs which is an indication of a potential phishing attack. This feature is identified in the literature (Fette, Sadeh & Tomasic, 2007).

“HREF” attribute and LINK Text mismatch

A link to another website is usually defined by use of html <a> anchor tag. “href” attribute allows a user to visit another website by describing the location of the second website to be visited. The link is rendered to the browser after a user clicks the “link text” (e.g., a href=“URL Address”>Link text). Link text could be a plain text, image or any element. If there is a match between the link text and the pointed website, then the website could be a phishing website. All the emails are checked for mismatch between the link text and the href attribute recording a positive Boolean feature if found.

Availability of “Link,” “Click,” and “Here” in Link Text of a Link

Links in most phishing emails contain certain words like “Click,” “Here”, “Login”, or “Update”. All emails are checked for presence of these words and a Boolean value is recorded based on the presence or absence of these words.

Dot contained in domain name

According to Emigh (2007) (“Phishing attacks: Information flow and chokepoints”), the number of dots that should be contained in a legitimate domain name should not exceed three. If the number of dots in URL exceeds three, then a binary value of 1 is recorded to assist in phishing features.

HTML email

Every email is defined by MIME standards which defines the types of components contained in the email. The component content type which is defined by content-type attribute could be plain text denoted by “text/plain”, HTML denoted by “text/html”. According to Fette et al., an email is a potential phishing email if it contains “text/html” attribute. They argued that is hard to achieve phishing attack without using HTML links.

Use of JavaScript

JavaScript is a scripting language that is used to perform a particular action. This is accomplished by either embedding in the body of an email using <script > tags or in a link using anchor <a> tag. Some malicious actors use JavaScript to evade detection by hiding information from users with the use of JavaScript. Fette et al. if an email is found to contain a JavaScript code in either the body of the email or in a link then it is classified as potential phishing email.

Links found in an email

The number of links contained in an email is recorded and used as feature to detect phishing emails. According to (Yuan & Zhang, 2012), phishing emails normally contain multiple links to malicious websites thus multiple links should be used as a phishing detection feature.

Domain Names in an email

This refers to the number unique domain names extracted from all the referenced URLs. The occurrences are recorded and the value is used as feature. Each occurring domain name is counted once and any subsequent occurrence is discarded. It is believed if an email contains multiple domain names then it is a potential phishing email.

Body_From Domain Match

All the domain names contained in an email are extracted which are then matched with the sender's domain name. The senders' domain name if extracted from the "From" field of the email. If there is a mismatch between the comparison, this could be a potential phishing email.

Word List

Phishing emails usually contain some occurring words which can be used as phishing detection features. These words will be categorized into six different categories whereby each category will be used as a single detection feature. This translates to having six different phishing features. In each category, every word is counted and duplicates discarded (normalized). These categories are:

- i. Confirm; Update
- ii. Customer; Client; User
- iii. Restrict; Hold; Supesnd
- iv. Account; Verify; Notification
- v. Password; Click; Login; Username
- vi. Social Security; SSN

3.3 Training, testing and validation

3.3.1 Training module

The Training module includes three components of: "Input Matrix", "Target Matrix", and "Fitness Network" as follows.

- "Input Matrix": this matrix contains all the benign emails from Spamcorpus dataset and all the phishing emails from PhishCorpus dataset that the NN model uses in training stage. These emails have been already: parsed by the "Email Parser", sanitised by the "Email Sanitiser", and vectorised by the "Email Vectoriser", Figure.1. In our implementation, this matrix is a logical $14,370 \times 5$ matrix which represents a matrix with $14,370$ rows and 5 columns. $14,370$ represent the total number of the emails in our implementation, which is $6,656$ for benign emails and $7,714$ for phishing emails precisely. 5 represents the size of the assigned vectors to the emails which carries five features for each email: the number of links in the email body, whether or not the email is an HTML email, whether or not there is JavaScript in the email, the number of the email's parts, and the vector average.
- "Target Matrix": this matrix includes all the decisions (benign or phishing) for all the emails.

These decisions are for each and every email stored in the “Input Matrix”. In our implementation, this matrix is a logical $14,370 \times 1$ matrix where $14,370$ represent the total number of the emails while 1 represents the size of the assigned decision vector to each email which either carries 0 (benign) or 1(phishing) as a value.

- “Fitness Network”: this is the NN model with n layers with x inputs and y outputs where the data from ‘Input’ and ‘Target’ matrixes are used for training, validation, and testing, respectively. In our implementation, our NN model has 10 hidden nodes or 10 layers/neurons where 70% of the data from ‘Input’ and ‘Target’ matrices are used for training, 15% for validation, and 15% for testing.

3.3.2 Validation and Testing modules

The Validation and Testing modules of the NN model includes two components of “Sample Matrix” and “Output Matrix” as follows.

- “Sample Matrix”: this matrix contains sample data from the “Input Matrix”. The trained NN model uses the data in the “Sample Matrix” as inputs during the testing phase. In our implementation, this matrix is a logical $n \times 5$ matrix contains n sample data from the

“Input Matrix”.

- “Output Matrix”: this matrix contains output data for the data in the “Sample Matrix”. The trained NN model predicts the output values for the “Sample Matrix” and stores them in the “Output Matrix”. In our implementation, this matrix is a logical $n \times 1$ matrix contains output data for the emails represented in “Sample Matrix”. The trained NN model predicts the output value, in terms of an email being benign or phishing, for each email in the “Sample Matrix”. These predictions will be stored in the “Output Matrix” and will be used to evaluate the performance of the neural network. We used 70% of the entire dataset, which includes all the benign emails from PhishTank dataset and all the phishing emails from PhishCorpus dataset, for training, 15% for validation and 15% for testing. We used scikit learn framework to develop, train, validate and then test our classifier. model. Our developed NN model has 10 hidden layers, 5 input features, 1 output layer, and 1 output features, Figure 3. The captured results are discussed in the next section.

`S+3656+cikit-learn KFold` class will be used to automatically implement k-fold cross-validation on the given data set.

3.3.3 Data source

The data will be collected from two online sources, one for benign URLs and the other one for phishing URLs.

The benign URL dataset will be collected from Alexa; which is a free open source data repository site that ranks URLs based on their popularity and non-malicious. The phishing email will be retrieved from the PhishTank website which is a free community website that allows users globally to submit, verify, track and share phishing URL data (PhishTank, 2016).

The online datasets will be both cleansed by removing any duplicates and for the experiments both a training and testing set will be created. The training set consists of 4000 URLs, 3000 from the benign set and 1000 from the malicious set. The testing set consists of 7000 URLs, 3000 from the benign set and 4000 from the malicious set. All URLs were selected randomly, except any URLs selected in the testing set do not include those that were present in the training set.

The next step is to extract features from the URLs. To ensure quality between features, all numeric values will be normalised such that their values lie between 0 and 1. All features in the below table are counts and binary values of specific entities within the URL.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

References

- Abdehamid, N. (2015). Multi-label rules for phishing classification. *Applied Computing and Informatics*, Vol. 11 (1), 29-46.
- Abdelhamid, N., Thabtah, F., & Ayesh, A. (2014). Phishing detection based associative classification data mining. *Expert systems with Applications Journal*, 41(2014) 5948-5959.
- Abdelhamid, N., & Thabtah, F. (2014). Associative Classification Approaches: Review and Comparison. *Journal of Information and Knowledge Management (JIKM)*, 13(3).
- Aburrous, M., Hossain, M., Dahal, K. P., & Thabtah, F. (2010). Experimental case studies for investigating e-banking phishing techniques and attack strategies. *Journal of Cognitive Computation*, 2(3), 242-253.
- Afroz, S., & Greenstadt, R. (2011). PhishZoo: Detecting phishing websites by looking at them. In *Fifth International Conference on Semantic Computing* (18-21 September 2011). Palo Alto, California USA, IEEE.
- Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics*, vol. 2014, Article ID 425731, 6 pages.
- Altaher, A., Wan, T. C., & Almomani, A. (2012). Evolving fuzzy neural network for phishing emails detection. *Journal of Computer Science*, 8(7).
- APWG Phishing Attack Trends Reports (2018). <https://www.antiphishing.org/resources/apwg-reports/>.
- Basnet, R., Mukkamala, S., & Sung, A. H. (2008). Detection of phishing attacks: A machine learning approach. In *Soft Computing Applications Industry* (pp. 373-383). Berlin: Springer.
- Behdad, M., T. French, M. Bennamoun, & L. Barone (2012). Nature-inspired techniques in the context of fraud detection. In *IEEE Transactions on Systems, Man, and Cybernetics C*.
- Bouckaert, R. (2004). Bayesian network classifiers in Weka. In *Working paper series*. University of Waikato, Department of Computer Science. No. 14/2004. Hamilton, New Zealand.
- Bright, M. (2011) Miller Smiles [Online] Available at: <http://www.millersmiles.co.uk/> [Accessed 9 January 2016]. *Computer Engineering, and Applied Computing*, pp. 682-686.
- Brown, S., Ofoghi, B., Ma, L., & Watters, P. (2017). Detecting phishing emails using hybrid features. In *Symposia and workshops on ubiquitous, autonomic and trusted computing (UIC-ATC '17)*, IEEE, Australia.
- Cranor, L. F., J. I. Hong, & Y. Zhang (2016). Cantina: A content-based approach to detecting phishing web sites. In *16th International World Wide Web Conference (WWW '07)*, Canada.

- Cutler, A., & Breiman, L., (2007). *Random forests-classification description*. Department of Statistics Homepage.
- Emigh, A. (2007). Phishing attacks: Information flow and chokepoints. In *Phishing and countermeasures: Understanding the increasing problem of electronic identity theft*, USA.
- Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web* (pp. 649-656).
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.
- Gaines, B. R., & Compton, J. P. (1995). Induction of ripple-down rules applied to modelling large databases. *Intell. Inf. Syst.*, 5(3), 211-228.
- Gupta, M., P. Prakash, R. R. Kompella, & M. Kumar (2015). PhishNet: Predictive blacklisting to detect phishing attacks. In *IEEE Conference on Computer Communications*.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).
- Han, W., Cao, Y., & Le, Y. (2015). Anti-phishing based on automated individual white-list. In *4th ACM Workshop on Digital Identity Management (DIM)* (pp. 51-59). ACM USA.
- Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11, 63-90.
- Huber, M., Mulazzani, M., Leithner, M., Schrittwieser, S., Wondracek, G., & Weippl, E. (2011). Computer security applications. In *27th Annual Computer Security Applications Conference*.
- Khonji, M., A. Jones, & Y. Iraqi (2013). Phishing detection: A literature survey. *IEEE Communications & Surveys Tutorials*.
- Ledesma, R., Chou, N., Mitchell, J. C., & Teraguchi, Y. (2014). Client-side defence against web-based identity theft. In *11th Annual Network & Distributed System Security Symposium*, USA.
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill, New York, NY, USA.
- Mohammad, R., Thabtah F., & McCluskey L. (2015B). Phishing websites dataset. Available: <https://archive.ics.uci.edu/ml/datasets/Phishing>. Accessed: January 2016.
- Mohammad, R., Thabtah F., & McCluskey L., (2014A). Predicting phishing websites based on self-structuring neural network. *Journal of Neural Computing and Applications*, 25(2), 443-458.
- Mohammad, R., Thabtah F., & McCluskey L., (2015A). Tutorial and critical analysis of phishing websites methods. *Computer Science Review Journal*, 17, 1-24.
- Mohammad, R., Thabtah F., & McCluskey, L., (2014B). Intelligent rule based phishing websites classification. *Journal of Information Security* (2), 1-17. ISSN 17518709. IET.
- Mohammad, R. M., Thabtah, F. & McCluskey, L. (2013). Predicting phishing websites using neural network trained with back-propagation. In *World Congress in Computer Science, Computer Engineering, and Applied Computing* (pp. 682-686). Las Vegas.
- Nargundkar, S., Tiruthani, N., & Yu, W. D. (2017). PhishCatch—a phishing detection tool. In *33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC '17)*, USA.
- Nazif, M., B. Ryner, & C. Whittaker (2010). Large-scale automatic classification of phishing pages. In *17th Annual Network & Distributed System Security Symposium (NDSS '10)*. The Internet Society, USA.
- Platt J. (1998). Fast training of SVM using sequential optimization. In *Advances in kernel methods support vector learning* (pp. 185-208). MIT Press, Cambridge.
- Qabajeh I., Thabtah, F., & Chiclana, F. (2015). Dynamic classification rules data mining method. *Journal of Management Analytics*, 2(3), 233-253.

- Quinlan, J. (1993). *Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.
- Sadeh, N., Fette, I., & Tomasic, A. (2017). Learning to detect phishing emails. In *16th International World Wide Web Conference (WWW '17)*, Canada.
- Smadi, S., Aslam, N., Zhang, L., Alasem, R., & Hossain, M. A. (2015). Detection of phishing emails using data mining algorithms. *Computer and Information Sciences*, 1-8.
- Strobel, S., Glahn, S., Moens, M. F., & Bergholz, A. (2010). New filtering approaches for phishing email. *Journal of Computer Security*, 18(1), 7-35.
- Tan, C. L., Chiew, K. L., & Sze, S. N. (2017). Phishing webpage detection using weighted URL tokens for identity keywords retrieval. In Ibrahim, H., Iqbal, S., Teoh, S., & Mustafa, M. (Eds.), *9th International Conference on Robotic, Vision, Signal Processing and Power Applications*. Lecture Notes in Electrical Engineering, vol 398. Springer, Singapore.
- Thabtah F., Mohammad R., & McCluskey L. (2016B). A dynamic self-structuring neural network model to combat phishing. In the *Proceedings of the 2016 IEEE World Congress on Computational Intelligence*. Vancouver, Canada.
- Thabtah F., Qabajeh I., & Chiclana F. (2016A). Constrained dynamic rule induction learning. *Expert Systems with Applications*, 63, 74-85.
- Wattenhofer, R., Burri, N., & Albrecht, K. (2015). Spamato-an extendable spam filter system. In *Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS '15)*, USA.
- Witten, I. H., & Frank E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers.
- Yuan, Y., & Zhang, N. (2012). Phishing detection using neural network. <http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf>.
- Zhang, Y., Cranor, L. F., Hong, J. I., & Egelman, S. (2016). Phishing detection: An evaluation of anti-phishing toolbars. In *14th Annual Network & Distributed System Security Symposium*, USA.





Managing the Implementation of Information Technology in Schools

Alaa Sarsour & Raed Sarsour

West Timisoara University, Timisoara, ROMANIA

Received: 18 June 2023 ▪ Revised: 23 October 2023 ▪ Accepted: 16 November 2023

Abstract

Modern technology has had a transformative impact on educational systems, learning, and teaching practices. The advent of the internet and the development of various digital tools have opened up new opportunities for education. Unfortunately, teachers and scholars, who may have educated themselves or being educated in a whole different era and time of schooling, find it really difficult to teach in the new modern ways. Different sides often blame each other for not being ready to innovations: the public blames teachers and superintendents for not adopting new technology in school and in turn teachers blame the state itself for not providing enough money and time. The authors claim that steps must be taken to implement IT in schools, and list possible steps.

Keywords: information technology (IT), educational systems, implementation at school, obstacles, possible solutions.

1. Introduction

Modern technology has had a transformative impact on educational systems, learning, and teaching practices. The advent of the internet and the development of various digital tools have opened up new opportunities for education. Websites, online platforms, and educational apps have provided students with access to a vast amount of information and resources, allowing them to explore and expand their knowledge beyond traditional classroom boundaries. This access to information has not only facilitated academic learning but also encouraged self-directed learning and personal exploration.

In fact, students today have to search the whole country to try to find a single school that does not have technology usage in their educational systems. IT has been wildly spreading for its huge benefits of connecting the student and school even at home. The authors believe that information technology (IT) has many uses for teaching and learning and making a connection between the student and teacher. In fact, it can be a lifestyle and considered too valuable to not be used in our lives. But the essence of information technology in schools and colleges assumes that teachers should know how to use it as effectively as possible in favor of their students.

2. IT role in our lives

The world around us is in a state of fast and rapid changes and discovering new technologies. The pressure on schools has never been greater, as they have to keep up with all the changes and developments to reach a state where schools will be using the latest updates and technologies of IT (Lloyd, 2020). American public schools face challenges in adapting to the constant changes in the modern world. While it would be unfair to generalize and say that all American public schools have not been creative or have failed to keep up with technology, there are certainly areas where improvements can be made (Salam et al., 2018). One of the main factors contributing to these challenges is the lack of adequate funding for many public schools. Insufficient resources can limit the ability of schools to invest in innovative teaching methods and up-to-date technology. Additionally, the focus on standardized testing in the American education system has sometimes resulted in a more rigid and standardized approach to instruction, which may hinder creativity and individualized learning (Culp, Honey & Mandinach, 2005).

Indeed, educational technology has the potential to have a significant impact on educational systems. Over the past two decades, advancements in technology have opened up new possibilities for teaching and learning. When properly integrated into the educational process, technology can enhance student engagement, facilitate personalized learning experiences, and provide access to resources and information that were previously inaccessible. By leveraging educational technology, schools can create more interactive and engaging learning environments that cater to diverse student needs (Culp, Honey & Mandinach, 2005).

3. Modern schools' challenges and opportunities

Teachers and scholars, who may have educated themselves or being educated in a whole different era and time of schooling, find it really difficult to teach in the new modern ways. Teachers play a crucial role in leveraging educational technology effectively and engaging students in new and meaningful ways. It is essential for teachers to continuously update their skills and knowledge to keep up with the changing landscape of education. By embracing professional development opportunities, educators can learn how to integrate technology tools into their teaching practice, adapt instructional methods to meet the needs of diverse learners, and foster student engagement. They can explore innovative teaching strategies, collaborate with colleagues, and stay informed about the latest advancements in educational technology (Halili, 2019).

The old educational systems structured of a teacher having and giving all the knowledge students need in life as a whole and skills and knowledge they needed to master for their benefits and that's what educational systems revolved around for many decades (Halverson & Shapiro, 2012).

Nowadays, there are plenty of students who have switched to technology learning, unlike students who have access to technology just to acquire just in-time learning, is increasing dramatically all over the world, in spite of the lack of technology access at schools. The numbers of students that study using information technology is still massive despite not allowing phones or computer usage in school. According to Pew Internet & American Life Project study (Madden, Lenhart, Duggan, Cortesi & Gasser, 2013):

- 78% of teens now have a cell phone, and almost half (47%) of them own smartphones.
- 23% of teens have a tablet computer, a level comparable to the general adult population.
- 95% of teens use the Internet.
- 93% of teens have a computer or have access to one at home.

- 71% of teens with home computer access say the laptop or desktop they use most often is one they share with other family members.

In a study by the Pew Research Center (Purcell, Heaps, Buchanan & Friedrich, 2013) researchers found that:

- 56% of teachers of the lowest income students say that a lack of resources among students to access digital technologies is a “major challenge” to incorporating more digital tools into their teaching; only 21% of teachers of the highest income students report that problem.
- 49% of teachers of students living in low-income households say their school’s use of internet filters has a major impact on their teaching, compared with 24% of those who teach better off students who say that.

The internet and platforms like YouTube and Google have significantly expanded access to information and learning resources. These platforms offer a wealth of educational content that can be beneficial for both students and teachers. Embracing information technology can indeed enhance the learning experience and provide opportunities for more practical and effective acquisition of knowledge and skills.

For students, the internet provides access to a wide range of educational materials, including video tutorials, online courses, interactive simulations, and research databases. They can leverage these resources to supplement their classroom learning, explore their interests, and engage in self-directed learning. Teachers can also use technology to curate and share relevant resources with their students, create engaging multimedia presentations, and facilitate online discussions and collaborative projects.

Additionally, technology can help bridge the gap between theoretical knowledge and practical application. For example, virtual reality (VR) and augmented reality (AR) technologies can provide immersive experiences that simulate real-world scenarios, allowing students to apply their knowledge in a practical context (Anthes et al., 2016). Online platforms can facilitate virtual labs and simulations, enabling students to conduct experiments and engage in hands-on learning even if physical resources are limited.

Furthermore, technology can support professional development for teachers. Online courses, webinars, and educational communities enable teachers to enhance their subject knowledge, learn new instructional strategies, and connect with educators from around the world. Teachers can also leverage technology to streamline administrative tasks, track student progress, and personalize instruction based on individual student needs.

4. Obstacles and possible solutions for successful IT implementation at school

Sometimes, from the public point of view, teachers and managers are seen as the holdback and reason for not adopting new technology in schools. It is a common challenge faced in many educational systems. In the authors’ point of view, it is important to recognize that the adoption of new technology in schools involves multiple stakeholders, including teachers, managers, administrators, and policymakers. Blaming any single group for the lack of technology adoption oversimplifies the problem.

Different sides often blame each other for not being ready to innovations: the public blames teachers and superintendents for not adopting new technology in school and in turn teachers blame the state itself for not providing enough money and time. Sometimes they say the state is not even committed to apply modern technology in schools (Ware et al., 2016).

In addition, the lack of access to technology among low-income students can indeed hinder their ability to fully benefit from the advantages that modern technology offers in education. This can create disparities in learning outcomes and further perpetuate socioeconomic inequities (Prensky, 2010).

To address this issue, it is crucial for educational institutions and policymakers to prioritize bridging the digital divide. This can be done through various means, such as:

1. Increasing access to technology. Schools and districts can work towards providing devices, internet connectivity, and digital resources to students from economically disadvantaged backgrounds. Efforts can be made to secure funding or form partnerships to ensure that all students have equal access to technology.
2. Technology integration in schools. It is important for schools to incorporate technology into their instructional practices and curriculum. This can help students develop digital literacy skills and better prepare them for the demands of the modern world. Teachers can play a vital role in this process by receiving training and support to effectively use technology in their classrooms.
3. Collaboration and partnerships. Collaboration between schools, community organizations, and government agencies can help expand access to technology and bridge the digital divide. By working together, stakeholders can pool resources, share expertise, and develop initiatives that address the specific needs of low-income students.
4. Alternative educational models. It's worth exploring alternative educational models that leverage technology to provide quality education to students who may not have access to traditional schools. Online learning platforms, blended learning models, and virtual classrooms can help bridge the gap and provide flexible learning opportunities.

Efforts to address the digital gaps should be accompanied by a commitment to provide equal educational opportunities for all students, regardless of their socioeconomic backgrounds. By focusing on equity and access, we can strive to ensure that technology enhances education rather than exacerbates existing disparities.

5. Summary

The availability of information through the internet, search engines like Google, and platforms like YouTube has significantly expanded the possibilities for acquiring knowledge and skills. These resources offer a wealth of information, tutorials, and educational content that can benefit both students and teachers.

However, it's important to note that while information technology offers tremendous potential, it should be implemented thoughtfully and with a focus on digital literacy and critical thinking skills. Students and teachers need guidance in evaluating the quality and reliability of online information and understanding how to effectively use technology for learning purposes.

By embracing information technology and incorporating it into education, we can create a more dynamic and inclusive learning environment that prepares students for the challenges of the modern world.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

References

- Anthes, C., García-Hernández, R. J., Wiedemann, M., & Kranzlmüller, D. (2016, March). State of the art of virtual reality technology. In *2016 IEEE aerospace conference* (pp. 1-19). IEEE.
- Culp, K. M., Honey, M., & Mandinach, E. (2005). A retrospective on twenty years of education technology policy. *Journal of Educational Computing Research*, 32(3), 279-307.
- Halili, S. H. (2019). Technological advancements in education 4.0. *The Online Journal of Distance Education and e-Learning*, 7(1), 63-69.
- Halverson, R., & Shapiro, R. B. (2012). *Technologies for education and technologies for learners: How information technologies are (and should be) changing schools*. Wisconsin Center for Educational Research (WCER), Working Paper, 6.
- Lloyd, I. (2020). *Information technology law*. Oxford University Press, USA.
- Madden, M., Lenhart, A., Duggan, M., Cortesi, S., & Gasser, U. (2013). *Teens and technology 2013*. Retrieved from <https://www.pewresearch.org/internet/2013/03/13/teens-and-technology-2013/>.
- Prensky, M. R. (2010). *Teaching digital natives: Partnering for real learning*. Corwin Press.
- Purcell, K., Heaps, A., Buchanan, J., & Friedrich, L. (2013). *How teachers are using technology at home and in their classrooms*. Washington, DC: Pew Research Center's Internet & American Life Project.
- Salam, S., Zeng, J., Pathan, Z. H., Latif, Z., & Shaheen, A. (2018). Impediments to the integration of ICT in public schools of contemporary societies: A review of literature. *Journal of Information Processing Systems*, 14(1).
- Ware, P., Kern, R., & Warschauer, M. (2016). The development of digital literacies. In *Handbook of second and foreign language writing* (pp. 307-328). De Gruyter Mouton.



A Hybrid Model for Detecting Insurance Fraud Using K-Means and Support Vector Machine Algorithms

Brian Ndirangu Muthura & Abraham Matheka

Kenyatta University, School of Engineering and Technology, Nairobi, KENYA

Received: 8 August 2023 ▪ Revised: 4 October 2023 ▪ Accepted: 15 November 2023

Abstract

Private stakeholders and governments across the globe are striving to improve the quality and access of healthcare services to citizens. The need to improve healthcare services, coupled with the increase in social awareness and improvement of people's living standards, has seen an increase in medical policyholders in the insurance industry. Even so, the healthcare sector is grappled with increased costs every other year, leading to revision of premiums and increased costs for the policyholders. One of the main factors contributing to the increased costs is fraudulent claims raised by the service providers and the policyholders, leading to unprecedented risks and losses for insurance firms. The insurance industry has set up fraud detection and mitigation systems to mitigate losses brought about by fraudulent claims, which come in two flavors: rule-based systems and expert claims analysis. With rule-based systems, conditions such as missing details, location of the claim vis a vis the location of the policyholder, among other rules, are evaluated by systems to assess the validity of the claims. On the other hand, insurance firms rely on the human intervention of experts using statistical analyses and artificial rules to detect fraudulent claims. The rule-based and expert analysis methods fail to detect patterns or anomalies in claims, which is central to efficient fraud detection. Data mining and machine learning techniques are being leveraged to detect fraud. This automation presents enormous opportunities for identifying hidden patterns for further analysis by insurance firms. This research aims to analyze a hybrid approach to detect medical insurance fraud using both K-Means (unsupervised) and Support Vector Machines (supervised) machine learning algorithms.

Keywords: fraud detection, machine learning, K-Means, support vector machines, hybrid algorithms.

1. Introduction

Insurance fraud is second to tax fraud in the frequency of occurrence (Association of Certified Fraud Examiners, 2019). The nature of the insurance business makes it susceptible to fraud. Insurance firms mainly manage risk for the policyholders by pooling and generating large cashflows through insurance premiums to pay loss claims. Insurance fraud occurs when the insured attempts to profit from the insurer while failing to comply with the policy's contractual terms and conditions, creating damage and losses for the insurer, and can occur at any stage of the policy term (Association of Certified Fraud Examiners, 2019). The losses span the long-term (comprised of life insurance) and short-term (comprised of motor and health insurance) insurance policies.

The prevalence of insurance fraud is not localized in one country but spread globally. For instance, the Coalition Against Insurance Fraud estimates that over \$80 billion is lost yearly due to insurance fraud in the United States of America. The Association of British Insurance recorded 107 000 fraudulent insurance claims in the United Kingdom in 2019 worth over £1.2 billion and depicted a 5% increase from 2018 (Association of Kenya Insurers, 2021). The national health fund in France (CNAM) estimated that \$321.4 million was lost due to fraudulent schemes and claims. The inquiries revealed that health providers such as doctors and practitioners accounted for the highest percentage of fraudulent claims, 48%.

On the other hand, health institutions such as hospitals and clinics accounted for 31% of the fraudulent claims against 21% by the insured. In India, the estimated losses attributed to fraud amount to \$6 billion annually, close to 8.5% of the total premiums remitted. The South African Insurance Crime Bureau estimated that out of \$2.4 billion in insurance claims paid in 2019, \$497.86 million could have been for fraudulent claims, which account for about 20% of the total claims raised in the year. In Kenya, the Insurance Fraud Investigation Unit identified 83 insurance fraud cases worth close to KES 386.34 million (Association of Kenya Insurers, 2021). The increase in insurance fraud, coupled with colossal sums of money involved, has led to an increase in the cost of insurance. Moreover, these figures are only estimates of claims deemed to be fraudulent and may not necessarily represent the precise magnitude of losses incurred.

Fraud in the insurance industry directly impacts a company's bottom line (Association of Kenya Insurers, 2021). Over 5% of a company's revenue is estimated to be lost to fraud yearly (ACFE, 2019). Through proper fraud detection mechanisms and validation of claims, insurance firms stand to benefit from increased profitability. Apart from the loss of revenue for the insurance firms, fraud schemes lead to the loss of the reputation of the insurers (Association of Certified Fraud Examiners, 2019). Insurance fraud is a global issue affecting the economy, state, community, and individuals.

Detecting and preventing fraud is a critical concern in the insurance industry (Matloob & Khan, 2019). While expert experience is critical in determining if a claim is fraudulent, the number of claims significantly raised surpasses the few experts tasked with analyzing these claims making it challenging to examine all insurance claims in real-time (Hanafy & Ming, 2021). Furthermore, differing experiences and perspectives from experts while dealing with the same claim cases contribute to decision bias. In the medical insurance industry, fraud detection has shifted from traditional domain expert analysis to rule-based systems (Gupta et al., 2021). The rule-based system contains sets of conditions that evaluate the validity of a claim which is an improvement from the domain expert analysis as the throughput of claims analyzed is much higher. However, there is an underlying need for even more efficiency in detecting insurance fraud in medical insurance (Matloob & Khan, 2019).

Researchers have proposed machine learning as a sophisticated technology that can be harnessed to assess claim patterns in medical insurance claims (Matloob & Khan, 2019). Machine learning can be applied to large datasets to discover unknown patterns and predict outcomes vital in fraud detection (Rawte & Anuradha, 2015). In medical insurance fraud detection, supervised learning is used to solve the classification problem into predefined labels (fraudulent and legitimate claims). In contrast, unsupervised machine learning algorithms address the clustering problem mainly through outlier detection (Rawte & Anuradha, 2015).

2. Problem statement

The implementation of robust fraud detection mechanisms is critical for medical insurance firms in the fight against fraudulent practices in the industry. A practical approach to fraud detection presents the opportunity of mitigating the losses attributed to fraudulent claims.

Subsequently, this presents an opportunity to reduce the cost of private medical insurance for the policyholders and increase profitability for the firms (Hanafy & Ming, 2021). Insurance firms have relied on domain expert analysis to detect fraud despite the benefits of a robust fraud detection technique. More recently, this approach has been automated by rule-based systems, which evaluate the validity of claims based on a set of rules as defined by domain experts. However, these attempts are ineffective in addressing fraud detection in the healthcare industry (Gupta et al., 2021).

Recent studies in fraud detection using machine learning and data mining techniques have focused on the efficacy of exclusively implementing supervised or unsupervised models. The supervised algorithms are mainly used to classify claims based on predefined labels, genuine and fraudulent claims. Similarly, unsupervised algorithms are used in clustering and outlier detections and do not need predefined labels. The combination of both supervised and unsupervised machine learning algorithms is complementary. Supervised algorithms learn from past fraudulent patterns, while unsupervised techniques target detecting new fraud patterns (Carcillo et al., 2021). The hybrid learning approach to fraud detection combines the advantages of supervised and unsupervised learning algorithms while reducing the inherent risk associated with either algorithm (Bauder et al., 2017). There is minimal research on using hybrid machine learning algorithms in fraud detection, more so with the combination of SVM and K-Means to analyze and solve classification and clustering problems, respectively.

3. Literature review

The chapter discusses the types of fraud in the medical insurance industry, the advancement of fraud detection techniques, and the use of supervised, unsupervised, and hybrid machine learning algorithms in fraud detection and prevention.

4. Types of medical insurance fraud

Healthcare fraud can be categorized based on the servicing pattern, that is, service-availing and service-providing patterns. The service-availing patterns are defined as fraudulent activities undertaken by the insured, while the service-providing patterns refer to the misrepresentation by the medical professionals (Matloob et al., 2020).

Healthcare fraud can also be categorized into service provider fraud, insurance subscriber fraud, insurance provider fraud, and conspiracy fraud. Service provider fraud may consist of charges incurred for medical services not performed, overbilling by the service provider, unbundling one medical procedure to multiple treatment stages, and billing each stage separately rather than consolidating, falsifying patients' diagnosis, and treatment history to validate superfluous medical procedures. Policy subscribers may commit fraud by filing claims for not receiving medical services, falsifying onboarding details to obtain a lower premium rate, and illegally claiming insurance benefits using another policyholder's coverage. On the other hand, insurance providers may commit fraud if they misrepresent the benefit offered for a particular scheme or product or make fake reimbursements to the service providers and policyholders. In conspiracy fraud, a combination of more than one of the tripartite parties is involved in getting undue benefits (Waghade & Karandikar, 2018).

5. Health care fraud detection methods

Fraud detection in medical insurance has evolved over the years. Traditional fraud detection techniques relied on rule-based methods. Claims would be evaluated for fraud based on

the rules outlined by domain experts (Zhou & Zhang, 2020). The efficacy of the rule-based evaluation method was constrained by the correctness of the rules (Zhou, He, Yang, Chen & Zhang, 2020). Traditional rule-based fraud detection methods relied on a few auditors to handle thousands of claims (Waghade & Karandikar, 2018). Only experienced auditors were able to uncover fraudulent claims. This approach was inefficient and time-consuming.

Electronic claim management systems have recently been implemented and integrated with healthcare systems (Kose, Gokturk & Kilic, 2015). Claim management systems are increasingly harnessed for auditing, review, and automatic claim processing. Electronic Claim Processing systems offer more efficiency and higher claim analysis throughput than traditional expert domain analysis (Ai, Lieberthal, Skyla & Wojciechowski, 2018).

The advancements in artificial intelligence, machine learning, and deep learning have resulted in new automated fraud detection methods with data mining and regression as the main approaches in medical insurance fraud detection (Joudaki et al., 2015).

6. Related works

Segal (2016) provides an introductory analysis of the capabilities of data mining techniques in pattern identification and matching in large data sets. Data mining can provide insights and trends in the underlying models. Moreover, data mining tools can detect anomalies and outliers by comparing them with known models and profiles.

In their study, Bauder et al. (2017) evaluate the use of data mining techniques in fraud detection and prevention by medical insurance firms. They posit that data mining in healthcare fraud detection involves structured and unstructured data. Structured data is a standardized data format that can be stored in tabular format in conventional databases. On the other hand, unstructured data is unformatted and unorganized data. Structured data is easy to analyze and model using data mining algorithms, while unstructured data requires additional steps such as parsing to analyze.

Zhou and Zhang (2020) explore in detail how data mining is used for fraud detection in the healthcare industry. They assert that fraud in medical treatment can be detected by analyzing abnormal data records and converting the fraud detection question into the classical outlier detection problem. The outlier detection method in data mining can further be divided into statistic-based, distance, clustering, and classification. The classification anomaly detection problem divides datasets into normal and abnormal types. In this outlier detection approach, the labelled data is applied for training and converts anomaly detection into a two-classification problem. Distance-based anomaly detection evaluates the range between the data sets (Zhou & Zhang, 2020). The local outlier factor (LOF) can be applied to each dataset to gauge the distance for each dataset. A large LOF value increases the likelihood of the dataset being an outlier. The statistical-based anomaly detection method assumes that outlier datasets do not conform to the model's distribution law of normal data. In the cluster-based approach, normal points often belong to clusters with multiple data points, while outliers belong to clusters with fewer or no data points (Zhou & Zhang, 2020).

Lawand and Kulkarni (2019) analyze insurance fraud prediction by solving the classification problem of the input space. Their research harnesses the Random Forest, decision tree, Naïve Bayesian Classification, and SVM algorithms to build a robust model with increased accuracy. It is evaluated using recall and precision metrics derived from the confusion matrix.

Hanafy & Ming (2021) Compare 13 machine learning methods in fraud detection in the insurance industry to show the impact of imbalanced datasets on the accuracy of the analysis. Resampling techniques such as Random Over Sampler, hybrid methods, and Random Under

Sampler are implemented to address the imbalanced datasets, thus enhancing the performance of the machine learning algorithms. They conclude that classifier algorithms cannot make accurate predictions with imbalanced datasets. SVM performs best using the Random Over Sampler, while C5.0 performs best using SMOTE and Random Under Sampler.

In their research, Naik and Laxminarayana (2017) state that in SVM, each data item is plotted as a point in n-dimensional space, where n denotes the number of unique features, with the value of each feature relating to the value of a coordinate. Classification is performed by finding the hyperplane that distinguishes the classes. Moreover, SVM is robust in outlier detection.

Rawte and Anuradha (2015) developed a hybrid machine learning algorithm using Evolving Clustering Model and Support Vector Machine. The Evolving Clustering Model is chosen since claim data is dynamic and constantly generated. At the same time, the support vector machine is used to solve the classification problems, thus detecting outliers and duplicate claims. The downside is that other forms of medical fraud are not detected.

In their paper, Naik and Laxminarayana (2017) note that the K-Means learning algorithm solves clustering problems by classifying a given input space through several clusters (often denoted by k). The main concept is to define k centroids, each cluster having one centroid. These centroids should be placed carefully as different locations result in different outcomes. Data from the input dataset is associated with the nearest centroid. A loop is generated after every revision, and the k centroids may change their location until no further movements are possible.

Ogbuabor and Ugwoke (2018) compare the performance of K-Means and DBSCAN using Silhouette score values. The efficacy of the K-Means algorithm is evaluated using different distance metrics and different numbers of clusters. In contrast, the efficiency of the DBSCAN algorithm uses different distance metrics and the least number of points to form a cluster. The results indicate that K-Means and DBSCAN have solid inter-cluster separation and intra-cluster cohesion. Based on the research outcome, K-Means outperforms the DBSCAN algorithm in accuracy and execution time.

In their paper, Wakoli et al. (2014) apply the K-Means algorithm to medical claim records to cluster the claim type and the cost per claim. The Euclidean distance measure was used to flag suspicious claims that would be revalidated. Similarly, the research by Zhang et al. (2020) compares the fraud detection efficacy of clustering algorithms on known medical fraud records. From their research, traditional rule sorts had a 24% detection rate, while DBSCAN had a 33.0% accuracy. Similarly, K-Means, Isolation Forest, and Local Outlier Factor had a detection rate of 35.0%, 47.0%, and 45%, respectively.

7. Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was used for the research. Developed by a consortium of data mining agents through an initiative sponsored by the European Union, CRISP-DM depicted data mining as a six-phase cycle (Schröer, Kruse & Gómez, 2021). The methodology consisted of the following phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Ordering the phases in the CRISP-DM methodology is flexible (Schröer, Kruse & Gómez, 2021).

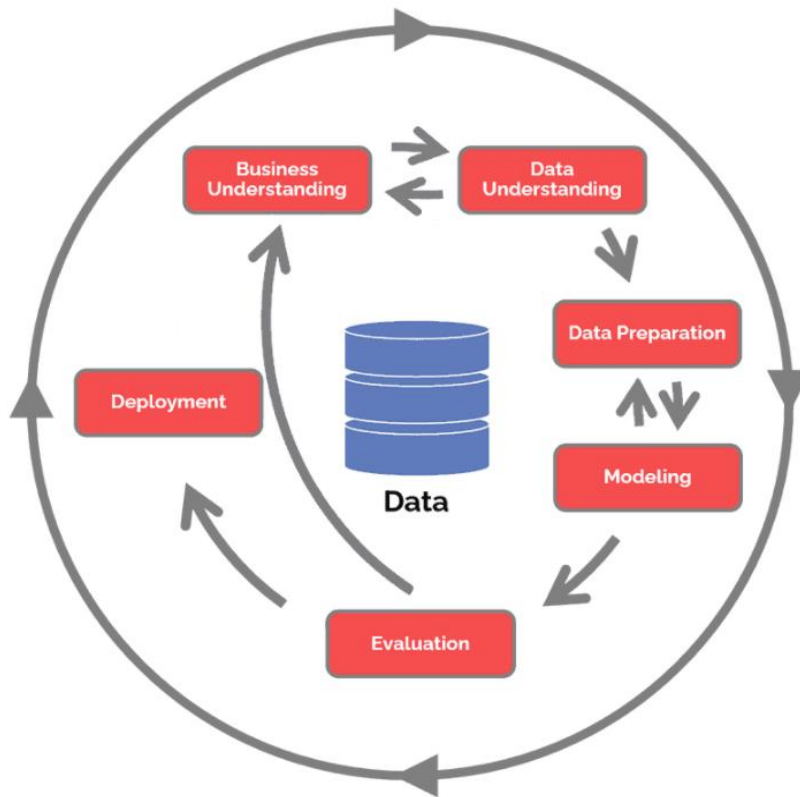


Figure 1. Phases of CRISP-DM (Schröer, Kruse & Gómez, 2021)

8. Business understanding

The business understanding phase of CRISP-DM involved a review of the journals and research papers on the various types of medical insurance fraud, how each fraud type is detected, and the current mitigation strategies for each fraud type assisting the research in gaining a more profound domain knowledge on the current controls, the evolution of fraud schemes, and which datasets can be mined for fraud detection in medical insurance. Moreover, this phase helped in selecting the appropriate technique to be used in answering the research questions.

9. Data understanding

The data understanding phase involved the collection of relevant medical claims datasets in tandem with the acquired domain knowledge. Subsequently, the research sought to familiarize with the claim datasets and ascertain the quality of datasets in developing the fraud detection model.

10. Data preparation

In this phase, the raw data provided was transformed into a standard acceptable format involving several activities, such as selecting relevant attributes and removing irrelevant attributes, handling null or missing values, and removing duplicate entries. The data cleansing step used a process called imputation which identified inaccurate and incomplete datasets,

substituted missing data with a placeholder, and noise reduction by removing data that did not relate to the research questions and objectives. Duplicated features that could be derived from existing feature sets or represented by another feature name were dropped. Identifying these features aided in identifying the strategy for feature engineering, feature relevance, and imputation strategies.

11. Modelling

The modelling phase attempted to solve the clustering and classification problems of the dataset involving implementing a hybrid machine learning approach where the K-Means algorithm was applied to the dataset for clustering similar features, and the Support Vector Machine (SVM) was harnessed for the classification of fraudulent and non-fraudulent claims. The merger of the two algorithms was achieved using a pipeline in Python.

12. Support vector machine model

The implementation phase sought to compare the performance of the hybrid fraud detection model vis the use of a sole supervised machine learning algorithm – SVM. These two models were also tuned as the last step of their iterations, and the performance metrics were recorded, resulting in four models: the lone SVM model, the tuned SVM model, the hybrid model, and the tuned hybrid model.

The transformed dataset was loaded with the SVM model with default hyperparameters for SVC, as shown in Table 1.

Table 1. Default Values for the SVC Model for Model 1

Parameter	Default Value
C	1
kernel	'rbf'
degree	3
gamma	'scale'
coef0	0
shrinking	TRUE
probability	FALSE
tol	1.00E-03
class_weight	None

The default hyperparameters were tuned using Python’s Grid Search Cross Validation library to obtain the optimal parameters for the SVM classification algorithm. These optimized and non-optimized predictions were later used as the benchmark for evaluating the classification performance of the hybrid model.

13. Hybrid machine learning model

The first iteration of implementing the hybrid model involved using a pipeline with the K-Means and SVM. The pipeline workflow was designed to run the standardized dataset with the default K-Means for clustering and classifying the output using SVM. The clustering and

classification were performed using the default kernel hyperparameters of the K-Means (shown in Table 2) and SVM algorithms (shown in Table 1), respectively.

Table 2. Default Values for the K-Means algorithm for model 3

Hyperparameter	Default Value
n_clusters	8
init	'K-Means++'
n_init	10
max_iter	300
tol	1.00E-04
precompute_distances	'auto'
verbose	0
random_state	None
copy_x	TRUE
algorithm	'auto'

The second iteration of the hybrid model applied the grid search library to exhaustively obtain the optimal parameters for the scaler, principal component analysis components, K-Means clusters, and the hyperparameter C in SVM. The parameter grid for the scaler parameter evaluated the Standard Scaler, Robust Scaler, and Quartile Transformer. The parameter grid for the PCA components ranged from 14 to 22 with an increment of 2. Similarly, the number of K-Means clusters ranged from 6 to 12 with an increment of 2.

14. Evaluation

After training the K-Means and SVM algorithms, the confusion matrix and the classification metrics were used to evaluate the efficiency and performance of the model on claims data on the insured. The model tested how many claims are categorized as false positives and false negatives (recall measure). Additionally, the model's performance was gauged by the percentage of correct classification of fraudulent claims (precision measure). The input space used a random resampling technique that rebalanced the imbalanced dataset's class distribution to improve the models' accuracy and reduce the false negatives and false positives. The performance evaluation metrics for the study included accuracy, precision, recall, and F1 score. These metrics were plotted on a Confusion Matrix to provide a detailed breakdown of the algorithm's true positive, true negative, false positive, and false negative predictions.

15. Results

The study summarized the performance of the four prototypes using a confusion matrix to evaluate the classification performance by categorizing the predicted and actual labels into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Classification accuracy, calculated by dividing the number of correct predictions by the total number of predictions, is a standard performance metric used to evaluate machine learning models by measuring the classified instances of true positives and true negatives (Altman & Krzywinski, 2017).

Table 3. Summarized accuracy of the models

Iteration	Model Description	Accuracy %
Model one	SVM with default hyperparameters	91.31%
Model two	SVM with optimal hyperparameters	97.05%
Model three	Hybrid model with default hyperparameters	68.08%
Model four	Hybrid model with tuned hyperparameters	97.49%

16. Confusion matrix

The study summarized the performance of the four prototypes using a confusion matrix to evaluate the classification performance by categorizing the predicted and actual labels into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

True Positives refer to the instances where the model predicted the positive class correctly with the actual label being positive. True Negatives occur when the model predicts a negative class correctly, thus matching with the negative label. False Positive (Type I error) occurs when the model incorrectly predicted a positive class, whereas the actual label is negative. False negative (Type II error) occurs when the model predicts a negative class, whereas the actual label is positive.

Table 4. Summary of the confusion matrix of the four models

Summary	TP	FP	FN	TN	Type I error rate	Type II error rate
Model one	1138	106	109	1122	4.28%	4.40%
Model two	1219	45	28	1183	1.82%	1.13%
Model three	784	328	463	900	13.25%	18.71%
Model four	1217	32	30	1196	1.29%	1.21%

The SVM model with default parameters noted a 4.28% Type I error and a 4.40% Type II error rate from the entire dataset. With the second iteration, both Type I and Type II error rates reduced to 1.82% and 1.13%, respectively, with an overall accuracy of 97.05% noted. The third prototype recorded increased Type I and Type II error rates at 13.25% and 18.75%, reducing the model's accuracy to 68%. The fourth model had the best accuracy rate of 97.45, mainly attributed to the lowest Type 1 error at 1.29%. The Type II error rate for the fourth model stood at 1.21%.

Further to the classification accuracy and confusion matrix, we examined other performance metrics such as precision, recall, and the F1 score to gauge the improvements of the four iterations of our model. Precision-measured the proportion of positively predicted cases against all predicted positive cases.

- Precision = $TP / (TP+FP)$

Recall measured the percentage of True Positives against all actual positive cases.

- Recall = $TP / (TP + FN)$

The F1 score evaluated the mean of precision and recall providing a balanced measure of the model's performance.

- F1 score = $2 * (precision * recall) / (precision + recall)$

Table 5. Summary of the classification report on the algorithms

	Precision:	Recall:	F-Score:
Model one	91.48%	91.26%	91.37%
Model two	96.44%	97.75%	97.09%
Model three	70.33%	62.55%	66.21%
Model four	97.44%	97.59%	97.52%

17. Discussion

While comparing the performance of the four prototypes, the hybrid model with optimized hyperparameters performed better than the other three prototypes in the classification of fraudulent claim transactions. Prototype 4 recorded the highest accuracy, precision, and f-score among the four models. However, prototype 2 recorded the highest recall score at 97.75%. Bauder et al. (2017) posit that hybrid machine learning models have the potential to outperform a single algorithm due to their robust nature, adaptability, complementary strengths, and fusion of decisions.

The introduction of hyperparameter tuning in model 1 and model 3 improved the accuracy of the base models. Hyperparameter tuning is selecting optimal parameters for a machine learning model. Hyperparameters control the behavior and influence the model's performance to find the best combination of parameters that will lead to the best performance of the model on a specific dataset. The study adopted the grid search approach, which predefined values for each parameter. Model 2 and Model 4 exhaustively evaluated all possible combinations of values defined in the grid set and returned the best parameter values for selection for each grid entry. Bergstra and Bengio (2012) note that grid search is computationally expensive for large search spaces and grid entries. In our study, the grid search hyperparameter tuning from Model 1 to Model 2 runs for approximately 305 seconds, while that from Model 3 to Model 4 runs for 3740 seconds.

The iteration from model three to model four introduced Principal Component Analysis to the pipeline before the k-Means algorithm step to reduce the number of dimensions in the dataset. Dimensionality reduction is the feature reduction process while preserving as much relevant information as possible to mitigate overfitting, noise reduction and improve the computational efficiency of the model. While there is no predefined cutoff for the number of components to be used in the PCA, Abdi and Williams (2010) suggest that the number of components selected should explain a high percentage of the total variance in the dataset. The introduction of PCA to model 4 increased the overall rise in performance metrics.

18. Model verdict

Based on the benchmark results of the SVM and in comparison, with the hybrid models, we note that the models with tuned hyperparameters scored better than those with the default parameters. Model 4 has the best accuracy, precision, and F1 scores in this case. Model 2 came in second with the best overall recall but second in accuracy, precision, and F1 scores. Model 1 was ranked third, with all performance measures being the third best. Model 3 was ranked in the fourth position. The hybrid classification model that uses both K-Means and SVM recorded a slight improvement in the classification of fraudulent and genuine claims compared to the classification of a single SVM model.

19. Conclusion

The research proposed, evaluated, and ranked the performance of a hybrid machine learning model that consisted of clustering using K-Means before classification using SVM. Various studies indicated hybrid machine-learning models perform better than a single algorithm (Zang & Ma, 2020; Bauder et al., 2017; Abdallah et al., 2017). This research adds to the existing knowledge base and elicits that hyperparameter tuning is a crucial step for performance metrics to be improved in hybrid algorithms. In as much as hyperparameter tuning adds to the model's accuracy, there needs to be consideration of its impact on the speed of the model's performance, especially if multiple steps are in the pipeline. Each parameter set for hyperparameter tuning increases the computation time exponentially. Nonetheless, a balance needs to be sought between improving the accuracy of the model vis a vis the acceptable execution time of the model.

20. Recommendation

The study's findings can be extended to the existing fraud detection models in the insurance industry with added accuracy by using singular classification algorithms. The study answers the question of the performance of the hybrid model in fraud detection. The study recommends an integrated approach with the model's prediction capabilities and core applications to detect fraud in real-time, which can be achieved using Application Programmable Interfaces (APIs) to get the classification rating based on the dataset's features.

21. Future research

While the developed model recorded an accuracy rate of 97.49%, further research needs to be conducted on improving the computational and speed performance of tuning hyperparameters in a hybrid machine-learning model. The study adopted grid search cross validation which exhaustively fits the parameter set. The study can be validated against other medical insurance firms to revalidate the outputs and reinforce learning.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

References

- Abdallah, A., Maarof, M., & Zainal, A. (2016). Fraud Detection System: A Survey. *Journal of Network and Computer Applications*, 90-113.
- Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.
- Ai, J., Lieberthal, R., Skyla, S., & Wojciechowski, R. (2018). Examining predictive modeling-based approaches to characterizing health care fraud. *Society of Actuaries*. <https://www.soa.org/resources/research-reports/2018/healthcare-fraud>.
- Altman, N. S., & Krzywinski, M. (2017). Points of significance: Classification evaluation. *Nature Methods*, 14(8), 755-756.

- Association of Certified Fraud Examiners (2019). *Insurance Fraud Handbook*. Association of Certified Fraud Examiners, Inc.
- Association of Kenya Insurers (2020). *2020 Insurance Industry Report*. Nairobi: Association of Kenya Insurers.
- Association of Kenya Insurers (2021). *Information Paper on Insurance Fraud*. Nairobi: Association of Kenya Insurers.
- Bauder, R., Khoshgoftaar, T., & Seliya, N. (2017). A Survey on the state of healthcare upcoding fraud analysis and detection. *Health Services & Outcomes Research*, 31-55.
- Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 281-305.
- Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Obleb, F., & Bontempi, G. (2021, May). Combining unsupervised and supervised learning in credit card fraud. *Business Analytics Emerging Trends and Challenges*, 557, 317-331.
- Gupta, R. Y., Mudigonda, S. S., & Baruah, P. K. (2021, March). A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage. *International Journal of Engineering Trends and Technology*, 96-102.
- Hanafy, M., & Ming, R. (2021). Using machine learning models to compare various resampling methods in predicting insurance fraud. *Journal of Theoretical and Applied Information Technology*, 99(12), 2819-2833.
- Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using data mining to detect health care fraud and abuse: A review of literature. *Global Journal of Health Science*, 194-202.
- Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing Journal*, 36, 283-299. <https://doi.org/10.1016/j.asoc.2015.07.018>
- Lawand, S., & Kulkarni, U. (2019). Survey on fraud prediction for an application using data mining. *International Journal of Emerging Technologies and Innovative Research*, 6(6), 209-212. <http://doi.org/10.1729/Journal.22988>
- Matloob, I., & Khan, S. (2019). A framework for fraud detection in government supported national healthcare programs. *Electronics, Computers and Artificial Intelligence, ECAI 2019*. Romania.
- Matloob, I., Khan, S., ur Rahman, H., & Hussain, F. (2020). Medical health benefits management system for real-time notification of fraud using historical medical records. *Applied Sciences*, 10(15). <https://doi.org/10.3390/app1015144>
- Naik, J., & Laxminarayana, A. (2017). Designing hybrid model for fraud detection in insurance. In *National Conference on Advances in Computational Biology, Communication, and Data Analytics*, 24-30.
- Ogbuabor, G., & Ugwoke, F. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal of Computer Science & Information Technology*, 10(2), 27-37.
- Rawte, V., & Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques. In *2015 International Conference on Communication, Information & Computer Technology (ICCICT)*.
- Schröer, C., Kruse, F., & Gómez, J. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 526-534.
- Segal, S. Y. (2016). Accounting frauds – Review of advanced technologies to detect and. *Economics and Business Review*, 45-64.
- Waghade, S. S., & Karandikar, A. (2018). A comprehensive study of healthcare fraud detection based on machine learning. *Nagpur: International Journal of Applied Engineering Research*.

Retrieved from https://www.ripublication.com/ijaer18/ijaerv13n6_140.pdf.

- Wakoli, L., Orto, A., & Mageto, S. (2014). Application of the K-means clustering algorithm in medical claims fraud / abuse algorithm in medical claims fraud / abuse detection. *International Journal of Application or Innovation in Engineering & Management*, 3(7), 142-151.
- Zhang, C., Xiao, X., & Wu, C. (2020). Medical fraud and abuse detection system based on machine learning. *International Journal of Environmental Research and Public Health*, 17(7265), 1-11.
- Zhang, Y., & Ma, S. (2020). *Ensemble machine learning: Methods and applications*. Springer.
- Zhou, S., & Zhang, R. (2020). A novel method for mining abnormal expenses in social medical insurance. *International IoT, Electronics, and Mechatronics Conference, Proceedings*. Institute of Electrical and Electronics Engineers Inc.
<https://doi.org/10.1109/IEMTRONICS51293.2020.9216354>
- Zhou, S., He, J., Yang, H., Chen, D., & Zhang, R. (2020). Big data-driven abnormal behavior detection in healthcare based on association rules. *IEEE Access*, 129002–129011.
<https://doi.org/10.1109/ACCESS.2020.3009006>



A Classifier Model to Detect Phishing Emails Using Ensemble Technique

Fredrick Nthurima & Abraham Matheka
Kenyatta University, Nairobi, KENYA
School of Engineering

Received: 11 September 2023 ▪ Revised: 18 November 2023 ▪ Accepted: 24 December 2023

Abstract

Phishing attacks usually take advantage of weaknesses in the way users behave. An attacker sends an email to the recipient that mimics a genuine email with phishing links. When the recipient clicks on the embedded links, the attacker can harvest critical information like credit card numbers, usernames or passwords as a result of entering the compromised account. Online surveys have put phishing attacks as the leading attack for web content, mostly targeting financial institutions. According to a survey conducted by Ponemon Institute LLC 2017, the loss due to phishing attacks is about \$1.5 billion annually. This is a global threat to information security, and it's on the rise due to IoT (Internet of Things) and thus requires a better phishing detection mechanism to mitigate these losses and reputation injury. This research paper explores and reports the use of multiple machine learning models by using an algorithm called Random Forest and using more phishing email features to improve the accuracy of phishing detection and prevention. This project will explore the existing phishing methods, investigate the effect of combining two machine learning algorithms to detect and prevent phishing attacks, design and develop a supervised classifier to detect and prevent phishing emails and test the model with existing data. A dataset consisting of benign and phishing emails will be used to conduct supervised learning by the model. Expected accuracy is 99.9%, with a rate of less than 0.1% for False Negatives (FN) and False Positives (FP).

Keywords: extractive model, abstractive model, hybrid model, natural language processing.

1. Introduction

Phishing attacks usually take advantage of weaknesses in the way users behave. An assailant directs an email to the recipient that mimics a genuine email with phishing links embedded in it. When the recipient clicks on the embedded links, the attacker can harvest critical information like credit card numbers, usernames or passwords as a result of entering the compromised account.

In this chapter, I will introduce the research problem and justification of the problem, what the research will achieve and address, research questions, research scope and the assumptions made during the research work.

Based on the Anti-Phishing Working Group Report 2018, a Phishing attack is the number one attack committed by threat actors as compared to other attacks. It is a form of fraud where the attacker deceives the target for personal gain or reputation damage. Fraud results in

users revealing their details like credit card numbers, passwords, PINs, usernames and other sensitive information leading to the compromise of accounts and loss of funds.

Phishing campaigns lure users into giving confidential information by visiting websites that look like legitimate ones (phishing.org, 2018). Phishing is done using a digital gadget like a computer or Ipad through a computer network. Malicious actors usually target the weakest element in the security chain, i.e., end-users (Khonj, Jones & Iraqi, 2013).

With a phishing attack, the attackers package messages so the target users cannot easily detect if the message is not genuine. The users end up clicking on the embedded links, thereby being redirected to the attacker's websites, whereby the attacker can get confidential information like passwords, usernames, credit card numbers etc. This enables threat actors to enter the compromised account and achieve their objectives like data theft, funds transfer or reputation injury.

For instance, a malicious email might have malware which, when clicked by the user, will install itself in the pc or mobile phone and will transfer funds to the account of the attacker whenever the owner of the account tries to transfer cash (Khonji et al., 2013). This attack is called Man in the Browser (MITB), a variant of the Man in the Middle (MITM) attack. The man-in-the-browser attack usually uses vectors like ActiveX components, plugins, or email attachments to deliver the payload to the user's computer or phone.

With the increasing case of cyber-attacks, organizations are looking for safer ways of protecting data and preventing getting hacked or getting hacked again. Design and technology should be greatly improved to prevent hackers from infiltrating networks.

According to (Behdad, French, Bennamoun & Barone, 2012), using better defense systems is not enough to stop malicious actors from penetrating systems since these are sometimes circumvented; a better system should detect malicious activities and prevent them before causing any damage.

1.1 Problem statement

Today, many spam email filters exist compared to filters for phishing emails. Many techniques are employed to develop phishing email filters, including Blacklists, Visual similarity, heuristics, and Machine Learning. The results of the above techniques have shown that Machine Learning does offer the best solution for phishing filters (Brown et al., 2017). However, current machine-learning anti-phishing solutions use a single model to detect phishing. According to the results, this could offer better detection accuracy, which currently stands at 98% (Smadi et al., 2015). Moreover, they have used domain/URL characteristics, leaving behind other phishing features in phishing emails and lowering accuracy and detection rates. There is a need to develop a better phishing classifier using a machine learning ensemble model, namely Random Forest, and include other phishing email features to increase detection accuracy. The Random Forest algorithm (RF) employed in this proposal work is a form of a bagging algorithm that categorizes many decision trees (from random training sets) to get improved classification accuracies (Deng et al., 2020).

2. Literature review

This chapter analyzes related works, techniques for text summarization, and various models in use.

2.1 Understanding phishing attacks

As per the Counter Phishing Working Gathering (APWG) report, the name “Phishing Movement Patterns Report – fourth Quarter 2017,” around 57% of phishing assaults target monetary foundations and settlement administrations. A phishing attack is a very common threat on the internet propagated by malicious actors who lure users into supplying personal information to their websites. By doing so, the malicious actors will be able to harvest critical information about a user ranging from passwords, credit card numbers or usernames, for their malicious objectives.

Researchers have demonstrated that social phishing, where in this case, the word *social* means information related to the target is used, produces very actual results as opposed to regular phishing. Gupta, Prakash, Kompella and Kumar (2015) concluded that if phishing attack emails mimicked a target's ally, the success rate of the phishing attack grew from 16% to 72%. Information's social aspect is valuable to social network operators and attackers. This is made even more possible if the information on social media contains an email address that is genuine or if there is a recent conversation between the target and the mimicked friend.

In the recent past, there has been an emergence with automation of data extraction from social media networks and sites. This has led to the availability of usable data to attackers, which can be used to carry out phishing attacks.

Ofoghi, Ma, Watters and Brown (2017) grouped the following spam attacks; Shared attribute attacks, Relationship-based attacks and Unshared attribute attacks.

With this kind of grouping, Relationship-based attacks use affiliation information only, making this spam attack look like socially engineered phishing which normally tricks users into clicking and inputting sensitive data. With the other attacks, they use information originating from social networks to compromise users and get sensitive data for their malicious actions.

This information originating from social networks is categorized between shared and unshared concerning the target and spoofed friend. Birthday cards can represent unshared information, which looks like genuine cards sent from the target's friend. On the other hand, common attributes like photos where both the victim and mimicked friends are both tagged can be abused for context-aware spam.

Huber, Mulazzani, Leithner, Schrittwieser, Wondracek and Weippl (2011) found that information from various social networks can be abused as a result of weaknesses in the communication channels. This will enable the attackers to acquire sensitive information for their gain. Furthermore, the authors have gone further to demonstrate that the data that is extracted from online networks can be exploited to aim many users with context-ware spam.

Gupta, Prakash, Kompella and Kumar (2015) used a hybrid of two techniques, namely blacklists and heuristics, to detect phishing emails. This hybrid technique attained a False Positive (FP) of 5% and a False Negative (FN) rate of 3%.

Holbrook, Kumaraguru, Downs, Cranor and Sheng (2010) researched several anti-phishing solutions and came up with ‘*SpoofGuard*’, which was designed by Ledesma, Chou, Mitchell and Teraguchi (2014). This solution ‘*SpoofGuard*’ showed an improved detection rate of 38% for False Positives (FP) and 9% for False negatives (FN). Moreover, Nargundkar, Tiruthani and Yu (2017) developed a phishing detection system that used heuristics as a mode of detection. This solution managed to achieve a False positive of 1% and a False Negative of 20%.

Smadi, Aslam, Zhang, Alasem and Hossain (2015) also used the heuristics technique, which achieved a False Positive rate of 3% and 11% False Negative.

Sadeh, Fette and Tomasic (2017) came up with a solution to detect phishing emails by use of Machine Learning. This technique achieved a False Positive rate of 1% and a False Negative

rate of 1.2%. Strobel, Glahn, Moens, De Beer and Bergholz (2010) came up with a hybrid solution by use of machine learning and heuristics, which achieved a False Positive rate of 0.05% and a False Negative rate of 1%.

The above techniques have relatively high False Positives and False Negatives. In our proposed anti-phishing technique, the features are extracted directly from the email, thus eliminating processing overhead and increasing run-time. Thus, by eliminating sending of queries, the proposed model will be faster and remove space complexities.

2.2 Machine learning anti-phishing methods

PhishHaven – An efficient real-time AI phishing URLs detection system

This constant computer-based intelligence-produced phishing URL arrangement of recognition was created by Maria Sameen, Kyunghyun Han and Seong Oun Hwang in 2020. This framework utilizes lexical elements-based extraction and investigation techniques. To expand the productivity of the framework, the framework utilizes URL HTML encoding as a lexical element. To detect tiny URLs, the system uses a URL hit mechanism. This system uses an ensemble machine learning model employing the multi-threading approach for the training and testing stages.

The framework utilizes fair democracy to allocate the last marks, i.e., typical or phishing, to the given URLs. This framework accomplished an exactness pace of 98% discovery. The framework involves a worldview execution for troupe AI, which includes equal execution of learning models through multi-stringing. Equal execution in the preparing and testing stages speeds up processes, consequently permitting the location of phishing URLs to progress. The proposed recognition framework flaunts different helpful highlights. First, it is free of any outsider administrations (i.e., WHOIS, Group Cymru, and so on) because every one of the methods, including highlight extraction from a URL assessment and characterization of a URL, is performed inside our location framework. Second, it is free of dialects since it dissects URLs, as it were. Furthermore, third, it can recognize zero-day assaults because the discovery framework dissects URLs in view of the URL’s lexical highlights.

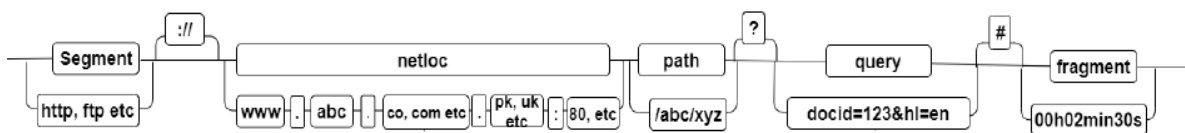


Figure 2.1. Lexical features approach from components of URL

2.3 Current application of phishing classifiers

Phishing emails exhibit different features that make them get distinguished from benign emails. These features include; subdomain, prefix_suffix, URL length etc. Mohammad, Thabtah and McCluskey (2013) created unique learning bases using space understanding to detect phishing and legitimate emails. Recent research shows on how to automate the detection of benign and phishing emails. The use of statistical analysis has been used to achieve this, according to Abdelhamid, Thabtah and Ayesh (2014). To study phishing emails better, emails from various sources were grouped based on various phishing features. This grouping was achieved through the recording of occurrences of phishing emails. To improve the detection rate, a larger dataset was collected from various sources (Abdelhamid, Thabtah & Ayesh, 2014).

Several methods have been used to study phishing patterns. These methods are decision tree, support vector machine, Random Forest and Naïve Bayes. A solution called PILFER, which stands for “*Phishing Identification by Learning on Features of Email Received*,” was designed to help curb the phishing menace. This solution was used with a case study of 860 phishing emails and 695 benign emails. This experiment was conducted to determine the phishing features in the emails. The features detected by this solution and experiment include IP-based URLs, Email body in HTML format, presence of JavaScript, number of links inside the email and others. Therefore, it was found that PILFER is good at improving the detection of phishing emails by considering the features found in the emails.

A method called the Random Forest algorithm was used against 2,000 email messages. This experiment aimed to reduce false positives and false negative rates (Akinyelu & Adewumi, 2014). When Random Forest is used with a combination of 15 features, it registers a significant reduction in error rate, becoming the best method in phishing classification and detection hence fitting. Phishing detection models using Random Forest are more dominant concerning detection rate.

Aburrous, Hossain, Dahal and Thabtah (2010) used identified features to classify websites by accurately classifying the identified features. The manual classification was used to group these features into six categories. The categories were then loaded into Waikato Environment for Knowledge Analysis (WEKA) for analysis. This analysis used instances totaling 1006 from PhishTank, whereby four classification algorithms were used to run several experiments. The effectiveness of the features used was measured by classification accuracy. In the experiments using decision tree algorithms, the authors noted a detection rate of 83% of the phishing sites. The authors further pointed out that when this algorithm is coupled with pre-processing, detection accuracy is significantly improved and would be used to make a very good detection model.

A Machine Learning covering algorithm, which goes by the name, Enhanced Dynamic Rule Induction (eDRI), is among the first algorithms to be used as an anti-phishing solution (Thabtah, Qabajeh & Chiclana, 2016). To process the datasets, this Covering algorithm uses frequency and Rule strength as the two major starting points. eDRI only stores “strong” features of the datasets if their frequency exceeds the minimum frequency threshold after scanning all the presented datasets. The stored features are incorporated in the rule, whereas all other values are gotten rid-off in this first process. eDRI removes its training cases, and then strong feature occurrences are updated to signify the inexistence of the instances. This process is done when a rule has been realized. This means eDRI removes its instances and retains strong features. This means eDRI removes features by itself, providing better controllable phishing models. In order to determine eDRI reliability, experiments were carried out on multiple phishing websites. 11,000 websites were collected for these experiments. eDRI showed better results than decision trees and other covering algorithms regarding phishing detection rate. A technique called trial and error Neural Networks which uses Machine Learning, has been condemned due to its time consumption (Mohammad, Thabtah & McCluskey, 2013). For this technique to be effective, a person knowledgeable about the domains is needed during the tuning phase. The elimination of trial and error was proposed but adopted a better self-structuring classification (*Ibid.*, 2016). The authors improved the phishing model by improving the learning rate and other parameters and later adding new neurons to the layer that is not visible. This means the features used to build the model are updated during the process of classifier model design.

According to Mohammad, Thabtah and McCluskey (2015), using a dynamic Neural Network model aimed to identify phishing cases from the dataset. Different dataset sizes were used to achieve this, i.e., 100, 200, 500, and 1,000. These experiments showed improved predictions in comparison to Bayesian networks and decision tree techniques.

Since phishing attackers constantly update their phishing techniques, there was a need to develop a more resilient model based on the previous training results (Thabtah, Mohammad & McCluskey, 2014). This aimed to develop a self-learning model to counter the ever-changing techniques used by phishers. The above Neural network algorithm tracks the model's performance by using smart decisions on the results of the validation dataset. The training phase goes as follows; when the error is below the minimum, the algorithm saves up the weights and proceeds with the process. However, if the fault exceeds the lower limit, the algorithm goes further without saving any weights. Parameters can be frequently updated without waiting for the model to be completely built. This experiment revealed that the Neural network model resulted in superior prediction rates compared to traditional techniques like C4.5 and probabilistic.

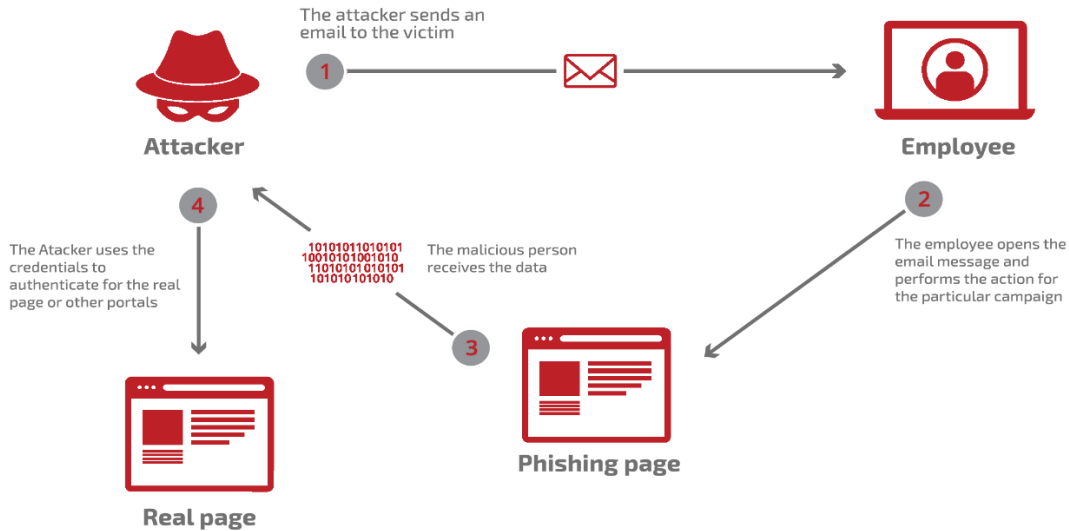


Figure 2.2. Phishing attack execution

2.3.1 Features used

This section describes the phishing features that our classifier will use. The features were extracted and identified from the literature and will form a combination of features that effectively classify phishing and benign emails. In this project, we will use 15 features identified from different literature commonly used by phishing attackers.

2.3.2 IP-based URLs

Legitimate websites usually have their names on the URL. A case like <http://www.mytours.com/> informs the user that someone will visit a website with the domain *mytours.com*. Attackers usually mask their identity by replacing the domain name with an IP address, e.g., <http://42.56.100.21/login.asp>. By doing this, malicious actors can escape detection by using IP-based URLs, which indicates a possible phishing attack. This discussed feature is identified in the literature (Fette, Sadeh & Tomasic, 2017).

2.3.3 LINK text mismatch and “HREF” attribute

A link to another website is usually defined by using an HTML <a> anchor tag. “href” attribute allows a user to visit another website by describing the location of the second website. The content is displayed on the browser when the user clicks the link. This link is in the form of a

href="URL Address"> link text . The link text can be plain text, an image or any element. If there is a match between the link text and the pointed website, the website could be phishing. Two items are checked for mismatch, i.e., link text and href attribute for all the emails. A positive Boolean is recorded when a mismatch is found on these emails.

2.3.4 *Link text of hyperlink*

Phishing emails exhibit certain characteristics on the links that make the emails qualify to be phishing emails. The emails will contain certain words like *click here*, *log in* or *update*. Emails are checked for the presence of these words, and a Boolean value is recorded if these words are found or not.

2.3.5 *Dot contained in domain name*

According to Emigh (2016), a legitimate domain name should contain less than three dots. If the number of dots in the URL exceeds three, a binary value of 1 is noted to assist in phishing features.

2.3.6 *HTML email*

MIME standards define every email. MIME standards define what makes up the email and its components. The components are categorized into two types, i.e., *text/plain* and *text/html*. These are the content-type according to Fette, Sadeh and Tomasic (2017), an email could be a phishing email if it has a "text/html" property. They argued that using HTML links is easier to achieve phishing attacks.

2.3.7 *Use of JavaScript*

JavaScript is a scripting language that is used to perform a particular action. JavaScript is either used in the body of the email using special tags denoted by <script> or can be used on a link using a tag called anchor <a>. Malicious actors make use of JavaScript language to evade detection by hiding information from users with the use of JavaScript. If an email contains a JavaScript code, it is classified as a potential phishing email (Fette, Sadeh & Tomasic, 2017).

2.3.8 *Links found in an email*

The sum of links in an email is registered to detect phishing emails. An email containing many links is a probable candidate for a phishing email. Phishing emails usually have links to external websites that redirect users to the attackers' websites (Yuan & Zhang, 2012).

2.3.9 *Email domain names*

The sum of unique domain characters is extracted for comparison with the referenced URLs. The incidences are recorded, and the value is used as a feature for detecting a phishing email. Each occurring unique domain name is recorded once, and any subsequent occurrence is discarded. It is therefore believed that if an email contains multiple domain names, it is a potential phishing email.

2.3.10 Body-from domain match

Domain names form a crucial part of phishing detection. This is because the domain identity of the sender and those in the body of the email should match if an email is to be classified as genuine. A match is performed on the sender’s domain name and that of the extracted domain names from the email. The “From” field gives the sender’s domain name and is compared with our test dataset for a match. If there is a disparity between the comparisons, this suggests it could be a potential phishing email (Altaher, Wan & ALmomani, 2012).

2.3.11 Word list

Phishing emails usually contain some occurring words which can be used as phishing detection features. These words will be categorized into six categories, each of which will be used as a single detection feature. This translates to having six different phishing features. Every word is counted in each category, and duplicates are discarded (normalized). These categories are:

- a) Confirm; Update
- b) Customer; Client; User
- c) Restrict, Suspend, Hold
- d) Notification, Account, Verify
- e) Password, Click, Username, Login
- f) Social Security; SSN

2.3.12 Email datasets, email classifier, email parser, email sanitizer, and email vectorizer ensemble model

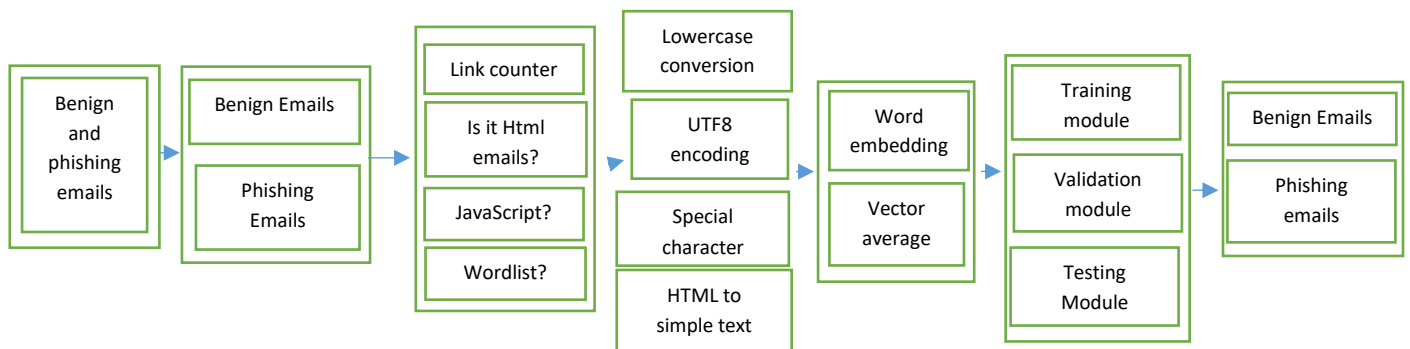


Figure 2.3. Proposed classifier model

3. Methodology

Machine learning is made up of the training phase and the testing phases. We intend to use two datasets to train our model; benign and phishing emails. We will obtain the dataset sets from Alexa for genuine (Benign) emails and PhishTank for luring (phishing) emails. We intend to use a combination of models through the use of an ensemble model called Random Forest (RF) to increase detection accuracy. EMBER (Open-source threat data training set), IBM Watson, Existing Anti-Phishing solutions like Spam Assassin and PhishTank and Alexa for data sets are tools to assist in data modeling. We will use Python libraries like sci-kit-learn, pandas, numpy, matplotlib, and Java.

This research will use a quantitative research method to answer the question of the model's accuracy.

3.1 Training, testing and validation

To train and test our classifier, we will use a method called 10-fold cross-validation. In this method, the training dataset will be prepared by classifying the dataset into 10 parts. Out of the 10 parts, 9 will be used to train our classifier, and the results obtained from this training will be used to validate the 10th group of the dataset. The process is repeated 10 times so that all ten parts will be used as training and testing data. The cross-checking technique ensures that the information used for training and testing are very different. In Machine Learning projects, this method of 10-fold cross-validation has proven to produce a very good error estimate of the classifier model.

Training the module

Regarding the training module, three constituents are involved: Input Matrix, Target Matrix and Fitness Network. These three components are used consecutively to train the classifier model better and increase the detection rate.

Input Matrix

At this stage, the model uses genuine emails from the Alexa dataset and phishing emails from PhishTank during the training stage of development. The first stage with these email datasets is to *parse* the emails by *email parser*. Then the emails are sanitized by what is known as *email sanitizer*; lastly, the emails are vectorized by what is known as *email vectorizer*. This research will have $x*5$ as the logical matrix, indicating 10,000 rows, and the other part of the matrix is 5, meaning 5 columns. 10,000 means there will be a total of 10,000 emails dataset, with 4000 being benign and the other part of 6,000 being known phishing emails.

Every email will have fifteen features with a vector size of 15.

Target matrix

At this stage, the decisions for all benign and phishing emails are found here. The emails stored in the input matrix each produce decisions found in this matrix. In this project, we will have a $10,000*1$ matrix meaning that 10,000 will be the total number of emails, whereas 1 will be vector size. The emails carry 0 or 1, where 0 denotes a benign email while 1 represents a phishing email.

Fitness network

This is where model formalization and testing takes place. The input and target matrix data are utilized in training, formalizing and testing. In this project, 15% will be used for validation, 15% for testing and 70% for training.

3.2 Model validation and testing

The validation and testing are the last stage in the model development. At this stage, two matrixes are used: Sample and output.

Sample matrix

This has data from the input matrix, which is usually sample data. After the model is trained, it uses data from the sample matrix, which is used during the testing stage. In our project, this matrix is an $m*5$ matrix containing sample data from the input matrix.

Output matrix

Data from the sample matrix produces data that is found in this matrix. After training the model, it stores output values in the out matrix. This project represents this by an $n * 1$ matrix which contains output data for emails represented in the sample matrix. Using the emails in the sample matrix, the trained model will predict if an email is benign or phishing. The output matrix will store these predictions and will be used to evaluate the performance of the Random Forest algorithm. To achieve our objectives, we plan to use the scikit learn framework to develop, train, validate and then test our classifier model.

The scikit-learn KFold class will automatically implement k-fold cross-validation on the given data set. We intend to use 10-fold cross-validation.

3.3 Data source

The experimental data will be collected from two different online sources, whereby one dataset will contain benign URLs while the other will contain phishing URLs. To collect data for the benign URL dataset will be collected from Alexa, which is a free, open-source data repository site that ranks URLs based on their popularity and non-malicious. The phishing email will be retrieved from the PhishTank website repository. This is a free community website that enables users all over the world to submit, confirm, analyze and share phishing URL data (PhishTank, 2016). The testing datasets will be prepared for testing by cleansing and ensuring no duplicates. This results in clean training and testing datasets. After the dataset preparation, the training dataset will comprise 4,000 URLs, 3,000 from the benign dataset and 1,000 from the malicious set. Moreover, the testing dataset will consist of 6,000 URLs, 2,000 from the benign dataset and 4,000 from the malicious set. To realize the best results, all URLs will be picked randomly, apart from any URLs that will be selected in the testing dataset that don't contain the sets in the training set.

The next stage will be to extract various features from the URLs that have been prepared and cleaned. To realize quality among features, numeral values will be normalized to be between 0 and 1. In this regard, the features are counts and binary representing values of specific entities within the URL.

3.4 Data set

The current data set consists of 6,000 emails, with 3,000 of them being phishing emails sourced from Alexa and PhishTank, and 3,000 legitimate emails obtained from the Spam Assassin website (Apacheorg, 2016). This data was collected with the intent of providing a comprehensive overview of the current phishing landscape and offer a basis for further data mining. The Spam Assassin has two different email types: those easily identified as legitimate and those that are hard to differentiate from spam. The hard-to-tell emails, while still legitimate, need a lot extra checking to ensure they are not actually spam.

Index	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	UsingIP	LongURL	ShortURL	Symbol@	Redirectin	PrefixSuffi	SubDomai	HTTPS	DomainRe	Favicon	NonStdPoi	HTTPSDon	RequestUF	AnchorUR	LinksInScri	ServerForr	
2	0	1	1	1	1	1	-1	0	1	-1	1	1	-1	1	0	-1	-1
3	1	1	0	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	-1	-1
4	2	1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	0	-1
5	3	1	0	-1	1	1	-1	1	1	-1	1	1	1	1	0	0	-1
6	4	-1	0	-1	1	-1	-1	1	1	-1	1	1	-1	1	0	0	-1
7	5	1	0	-1	1	1	-1	-1	-1	1	1	1	1	-1	-1	0	-1
8	6	1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	-1	-1
9	7	1	0	-1	1	1	-1	1	1	-1	1	1	-1	1	0	1	-1
10	8	1	1	-1	1	1	-1	-1	1	-1	1	1	1	1	0	1	-1
11	9	1	1	1	1	1	-1	0	1	1	1	1	1	-1	0	0	-1
12	10	1	1	-1	1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1
13	11	-1	1	-1	1	-1	-1	0	0	1	1	1	-1	-1	-1	1	-1
14	12	1	1	-1	1	1	-1	0	-1	1	1	1	1	-1	-1	-1	-1
15	13	1	1	-1	1	1	1	-1	1	-1	1	1	-1	1	0	1	1
16	14	1	-1	-1	-1	1	-1	0	0	1	1	1	1	-1	-1	0	-1
17	15	1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	0	-1	-1
18	16	1	-1	1	1	1	-1	-1	0	1	1	-1	1	1	0	-1	-1
19	17	1	1	1	1	1	-1	-1	1	1	1	1	-1	-1	0	-1	-1
20	18	1	1	1	1	1	-1	-1	1	-1	1	1	1	1	0	0	-1
21	19	1	0	-1	1	1	-1	0	1	-1	1	1	1	1	0	0	-1
22	20	1	0	1	1	1	-1	0	1	1	1	1	-1	-1	0	-1	-1
23	21	1	1	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	0	-1
24	22	1	1	1	1	1	-1	1	0	-1	1	1	1	1	0	0	-1
25	23	1	-1	-1	-1	1	-1	1	1	-1	1	1	-1	-1	0	0	-1
26	24	1	-1	1	1	1	-1	0	1	-1	1	1	1	1	1	0	-1
27	25	1	-1	1	1	1	-1	0	-1	1	1	1	-1	-1	-1	-1	-1
28	26	1	-1	-1	1	1	1	-1	1	1	1	1	1	-1	1	0	-1
29	27	1	-1	-1	1	-1	1	-1	1	-1	1	1	1	1	0	-1	-1

Figure 3.3. Sample of the dataset 15 feature

4. Dataset split for training and testing

The evaluation of the performance of the classifying the phishing websites is demonstrated in Figure 4.4 through the use of three distinct machine learning algorithms: Random Forest, Decision Tree classifier and Adaboost. This provides a comprehensive analysis of the Classification Accuracy (CA) of each model in terms of effectiveness and efficiency.

4.1 Tool

To evaluate the compatibility of the file, it was converted to a CSV format and tested using the five algorithms selected by the WEKA tool. The results of this experiment will determine whether the file can be used with the WEKA tool or not.

Weka is a powerful set of machine learning algorithms designed to tackle a variety of data mining tasks. It provides a range of tools to pre-process, classify, regress, visualize, and cluster data. These algorithms can be used directly on the dataset or called from Java code, making it ideal for developing new machine learning approaches. With its perplexing capabilities and high burstiness.

Weka is a powerful tool for data mining, offering a broad variety of algorithms to help with any data mining task. This software provides users with the ability to analyze and uncover hidden patterns in large datasets. With Weka, users can quickly and easily explore, visualize, and manipulate data, and ultimately make more informed decisions (Weka, 2016).

4.2 Experimental results

Conduct various experiments in different scenarios, evaluate experiments and results using various measures, compare the performance of several experiments, and highlight the results.

The phishing classification model was implemented through generation of the Random Forest Classifier, Decision Tree Classifier and the AdaBoost algorithm. This project aims to evaluate the use of ensemble methods against other algorithms and determine which method is

better for ranking phishing and non-phishing websites. These algorithms are also generated using Python.

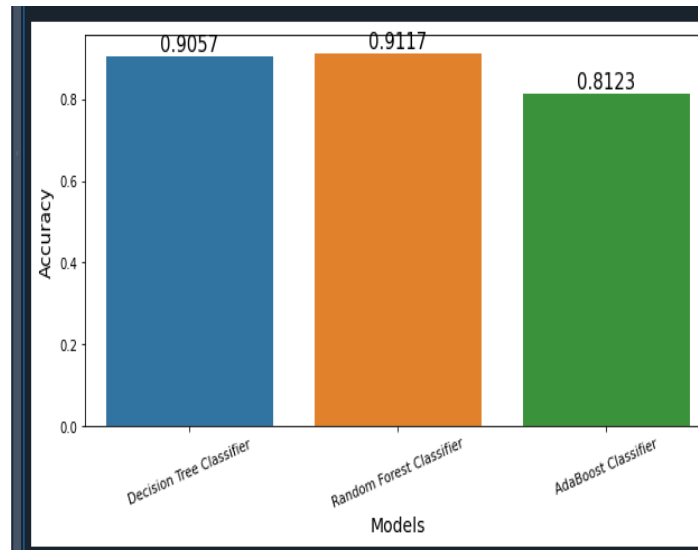


Figure 4.1 Bar plot accuracy of three algorithms

Table 1. Accuracy of three algorithms

No	Algorithm Name	Accuracy
1	Decision Tree	90.59%
2	Random Forest	91.15%
3	Adaboost	81.23%

Figure 4.6 and Table 1 illustrates a difference in algorithms and accuracy. Indicating that the Random Forest classifier gives the biggest accuracy, which is 91.15%, then the Decision Tree, 90.59%, and AdaBoost gives the smallest accuracy, 81.23%. This output demonstrates the accuracy percentage, whereas the training and testing sets are identified with different parameters in the dataset.

5. Conclusion and future work

5.1 Conclusion

The results of this project were based on expected results and were instrument tested. This chapter deals with the model’s contributions, limitations, suggestions and improvements. Contributions are collected through system goals. The challenges are evaluated by studying the completeness of the model or the challenges and difficulties encountered during the development process of the classifier.

Phishing emails are now a norm in recent years. Phishing is when the victim sends an email requesting key information from the user, which is sent directly to the phisher. Therefore, tracking these emails is necessary. There are numerous innovations to distinguish phishing messages. In any case, they all have restrictions, for example, low exactness, the substance might be like authentic messages and in this manner can’t be recognized, and the identification rate should be higher; thus, they have high bogus positives and high misleading negatives. This study evaluated the accuracy of phishing email detection through the use of manual selection feature and also the use of automatic feature selection of three classification algorithms that have high detection rates.

At the end of the process, the two scenarios are compared to determine the method that yields better results in terms of detection rates.

For manual attribute selection, 15 email attributes were chosen and divided into four categories based on email structure (body attributes, header attributes, URL attributes and Java script attributes with external attributes). The results indicate that the body group has the highest accuracy rate in detecting phishing emails, reaching 91.16%.

On the other hand, all but one of the four groups were tested together for accuracy each time.

The results indicate that the highest accuracy rate, 98.25, is achieved if the URL attribute group is removed from all the attributes.

Using auto-selection of the project testing, the accuracy was tested on three sets of auto-selected features, which are generated by the system. The results showed a deviation in accuracy between the three categories, with the highest group being the third one achieving 98% precision.

5.2 Future work

More work is needed for future feature selection techniques since selection techniques still need to be refined to cope with new techniques that anglers develop over time. Thusly, we propose to obtain another mechanized device to separate new elements from new crude messages to improve phishing email location precision and adapt to the extension of phishing methods.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

References

- Abdelhamid, N., & Thabtah, F. (2014). Associative classification approaches: Review and comparison. *Journal of Information and Knowledge Management (JIKM)*, 13(3).
- Aburrous, M., Hossain, M., Dahal, K. P., & Thabtah, F. (2010). Experimental case studies for investigating e-banking phishing techniques and attack strategies. *Journal of Cognitive Computation*, 2(3), 242-253.
- Afroz, S., & Greenstadt, R. (2011). PhishZoo: Detecting phishing websites by looking at them. In *Fifth International Conference on Semantic Computing* (18-21 September). Palo Alto, California USA, 2011. IEEE.
- Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing emails using random forest machine learning technique. *Journal of Applied Mathematics*, vol. 2014, Article ID 425731, 6 pages, 2014.
- Altaher, A., Wan, T. C., & Almomani, A., (2012). Evolving fuzzy neural network for phishing emails detection. *Journal of Computer Science*, 8(7).
- APWG Phishing Attack Trends Reports (2018). <https://www.antiphishing.org/resources/apwg-reports/>.
- Basnet, R., Mukkamala, S., & Sung, A. H. (2008). *Detection of phishing attacks: A machine learning approach*. Soft Computing Applications Industry, pp. 373-383.
- Bayesian network classifiers in Weka (2004). Working paper series. University of Waikato, Department of Computer Science. No. 14/2004. Hamilton, New Zealand: University of Waikato.
- Behdad, M., French, T., Bennamoun, T., & Barone, L. (2012). *Nature-inspired techniques in the context of fraud detection*. IEEE Transactions on Systems, Man, and Cybernetics C.
- Bouckaert, R. (2004). *Bayesian network classifiers in Weka* (Working paper series. University of Waikato, Department of Computer Science. No. 14/2004). Hamilton, New Zealand: University.
- Brown, S., Ofoghi, B., Ma, L., & Watters, P. (2017). Detecting phishing emails using hybrid features. *Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (UIC-ATC '17)*, IEEE, Australia.
- Cranor, L. F., J. I. Hong, & Y. Zhang (2016). Cantina: A content-based approach to detecting phishing websites. In *16th International World Wide Web Conference (WWW '07)*, Canada.
- Cutler, A., & Breiman, L. (2007). *Random forests-classification description*. Department of Statistics Homepage.
- Emigh, A. (2016). Phishing attacks: information flow and chokepoints. In *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. USA.
- Fette, I., Sadeh, N., & Tomasic, A. (2017). *Learning to detect phishing emails. Proceedings of the 16th international conference on the World Wide Web*. 649-656.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.
- Gaines, B. R., & Compton, J. P. (1995). Induction of ripple-down rules applied to modeling large databases. *Intell. Inf. Syst.*, 5(3), 211-228.
- Gupta, M., Prakash, P., Kompella, R. R., & Kumar, M. (2015). PhishNet: Predictive blacklisting to detect phishing attacks. *IEEE Conference on Computer Communications*.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1.

- Han, W., Cao, Y., & Le, Y. (2015). Anti-phishing based on automated individual white-list. *4th ACM workshop on digital identity management (DIM)* (pp. 51-59). ACM USA.
- Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63-90.
- Huber, M., Mulazzani, M., Leithner, M., Schrittwieser, S., Wondracek, G., & Weippl, E. (2011). Computer security applications. *27th Annual Computer Security Applications Conference*.
- Khonji, M, Jones, A., & Iraqi, Y. (2013). *Phishing detection: A literature survey*. IEEE Communications & Surveys Tutorials.
- Ledesma, R., Chou, N., Mitchell, J. C., & Teraguchi, Y. (2014). Client-side defense against web-based identity theft. *11th Annual Network & Distributed System Security Symposium*. USA.
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill, New York, NY, USA.
- Mohammad, R., Thabtah, F., & McCluskey L. (2015B). *Phishing websites dataset*. Available: <https://archive.ics.uci.edu/ml/datasets/Phishing+Websites>. Accessed January 2016.
- Mohammad, R., Thabtah F., & McCluskey L. (2014A). Predicting phishing websites based on self-structuring neural network. *Journal of Neural Computing and Applications*, *25*(2), 443-458. ISSN 0941-0643. Springer.
- Mohammad, R. M., Thabtah, F., & McCluskey, L. (2013). Predicting phishing websites using neural network trained with back-propagation. Las Vegas, *World Congress in Computer Science, Computer Engineering, and Applied Computing*, pp. 682-686.
- Nargundkar, S., Tiruthani, N., & Yu, W. D. (2017). PhishCatch – A phishing detection tool. *33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC '17)*, USA.
- Nazif, M., Ryner, B., & Whittaker, C. (2010). Large-scale automatic classification of phishing pages. *17th Annual Network & Distributed System Security Symposium (NDSS '10)*. The Internet Society, USA.
- Platt, J. (1998). *Fast training of SVM using sequential optimization: Advances in kernel methods support vector learning*. MIT Press, Cambridge, 1998, pp. 185-208
- Qabajeh I., Thabtah, F., & Chiclana, F. (2015). Dynamic classification rules data mining method. *Journal of Management Analytics*, *2*(3), 233-253.
- Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.
- Sadeh, N., Fette, I., & Tomasic, A. (2017). Learning to detect phishing emails. *16th International World Wide Web Conference (WWW '17)*. Canada.
- Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. *Proceedings of the 28th international conference on human factors in computing systems - CHI '10*, 373–382. <https://doi.org/10.1145/1753326.1753383>
- Smadi, S., Aslam, N., Zhang, L., Alasem, R., & Hossain, M. A. (2015). Detection of phishing emails using data mining algorithms. *9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*.
- Strobel, S., Glahn, S., Moens, M. F., & Bergholz, A. (2010). New filtering approaches for phishing email. *Journal of Computer Security*, *18*(1), 7-35.
- Sung, A. H., Basnet, R., & Mukkamala, S. (2008). Detection of phishing attacks: A machine learning approach. In *Soft Computing Applications in Industry*. Germany.
- Tan, C. L., Chiew, K. L., & Sze, S. N. (2017). Phishing webpage detection using weighted URL tokens for identity keywords retrieval. In Ibrahim, H., Iqbal, S., Teoh. S., & Mustafa, M. (Eds). *9th*

International conference on Robotic, Vision, Signal Processing and Power Applications. Lecture Notes in Electrical Engineering. Vol. 398. Springer, Singapore.

Thabtah, F., Mohammad, R., & McCluskey, L. (2016B). A dynamic self-structuring neural network model to combat phishing. In *Proceedings of the 2016 IEEE World Congress on Computational Intelligence*. Vancouver, Canada.

Thabtah, F., Qabajeh, I., & Chiclana, F. (2016A). Constrained dynamic rule induction learning. *Expert Systems with Applications*, 63, 74-85.

Wattenhofer, R., Burri, N., & Albrecht, K. (2015). Spamoto-an extendable spam filter system. In *Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS '15)*. USA.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. PubMed Central.

Yuan, Y., & Zhang, N. (2012). *Phishing detection using neural network*. <http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf>

Zhang, Y., Cranor, L. F., Hong, J. I, & Egelman, S. (2016). Finding phish: an evaluation of anti-phishing toolbars. *14th Annual Network & Distributed System Security Symposium*. USA.



AIMS AND SCOPE

The OJIT, as an international multi-disciplinary peer-reviewed **online open access academic journal**, publishes academic articles deal with different problems and topics in various areas of information technology and close scientific disciplines (information society, information communication technology - ICT, information architecture, knowledge organisation and management, information seeking, information management, electronic data processing – hardware and software, philosophy of information, communication theory and studies, mass communication, information ethics, library and information science, archival science, intellectual property, history of computer technology, development of digital competencies, ICT in education and learning, ICT education, etc.).

The OJIT provides a platform for the manuscripts from different areas of research, which may rest on the full spectrum of established methodologies, including theoretical discussions and empirical investigations. The manuscripts may represent a variety of theoretical perspectives and different methodological approaches.

The OJIT is already indexed in Crossref (DOI), BASE (Bielefeld Academic Search Engine), Google Scholar, J-Gate, ResearchBib and WorldCat - OCLC, and is applied for indexing in the other bases (Clarivate Analytics – SCIE, ESCI, and SCI, Scopus, Ulrich's Periodicals Directory, Cabell's Directory, SHERPA/RoMEO, EZB - Electronic Journals Library, etc.).

The authors of articles accepted for publishing in the OJIT should get the ORCID number (www.orcid.org).

The journal is now publishing 2 times a year.

PEER REVIEW POLICY

All manuscripts submitted for publishing in the OJIT are expected to be free from language errors and must be written and formatted strictly according to the latest edition of the [APA style](#). Manuscripts that are not entirely written according to APA style and/or do not reflect an expert use of the English language will **not** be considered for publication and will **not** be sent to the journal reviewers for evaluation. It is completely the author's responsibility to comply with the rules. We highly recommend that non-native speakers of English have manuscripts proofread by a copy editor before submission. However, proof of copy editing does *not* guarantee acceptance of a manuscript for publication in the OJIT.

The OJIT operates a double-blind peer reviewing process. The manuscript should not include authors' names, institutional affiliations, contact information. Also, authors' own works need to be blinded in the references (see the APA style). All submitted manuscripts are reviewed by the editors, and only those meeting the aims and scope of the journal will be sent for outside review. Each manuscript is reviewed by at least two reviewers.

The editors are doing their best to reduce the time that elapses between a paper's submission and publication in a regular issue. It is expected that the review and publication processes will be completed in about 2-3 months after submission depending on reviewers' feedback and the editors' final decision. If revisions are requested some changing and corrections then publication time becomes longer. At the end of the review process, accepted papers will be published on the journal's website.

OPEN ACCESS POLICY



The OJIT is an open access journal which means that all content is freely available without charge to the user or his/her institution. Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the articles, or use them for any other lawful purpose, without asking prior permission from the publisher or the author. This is in accordance with the BOAI definition of open access.



All articles published in the OJIT are licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Authors hold the copyrights of their own articles by acknowledging that their articles are originally published in the OJIT.

