International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 175-178, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1001

Real-Time Traffic Monitoring

Pooja S
Assistant Professor
Department of Electronics and Communication Engineering
K S Institute of Technology
Bangalore, India

C A Sushma
Department of Electronics and Communication Engineering
K S Institute of Technology
Bangalore, India

N Naga Omkar
Department of Electronics and Communication Engineering
K S Institute of Technology
Bangalore, India

Shiva Shankar B
Department of Electronics and Communication Engineering
K S Institute of Technology
Bangalore, India

Rithvik P
Department of Electronics and Communication Engineering
K S Institute of Technology
Bangalore, India

Abstract: As the population increases day by day vehicular travel is also increasing which leads to congestion problem. Traffic congestion causes many critical problems and challenges in the most populated cities. The increased traffic leads to more waiting time and fuel wastages. People miss opportunities, loose time and get frustrated. Traffic load is highly dependent on unpredictable situations such as accidents or constructional activities. These problems can be solved by a traffic control system by continuously sensing and adjusting traffic lights timing according to the actual traffic load which is called an Intelligent Traffic control System. The Intelligent Traffic Control Systems reduces congestion, operational costs, provides alternate routes to travelers and increases capacity of infrastructure.

Keywords: Traffic, Intelligent, Control, System, Alternate Route.

1. INTRODUCTION

Traffic investigate main aim is to optimize traffic flow of goods and citizens which causes lots of trouble especially when there are emergency case sat traffic light intersection which is always busy with lots of vehicles. However there are some restrictions in handling intelligent traffic control systems. The Density Based Signal executive in Traffic System is to solve traffic congestion difficulty which many people face and is a big problem in many cities. The system proposed here increases road safety even during the absence of traffic police and brings their attention to those who break the law. Traffic is coordinated in a circular loop that takes in the inputs in real time basics. NODE MCU with ESP Wi-Fi module is used to transfer and collect all the data from the sensors. All the data are made available at our local servers that are setup which will receive the data from the NODE MCU. The signals help at increasing the traffic-handling capacity at most intersections. They can function without any help from timers, connect to a computer controlled system which operates at few intersections.

2. LITERATURE SURVEY

W. Wen et al. [1] this paper proposed a framework for a dynamic and automatic traffic light control system and developed a simulation model to help design the system. The model adopts average departure times and arrival times which are physically observed at each intersection. Here by controlling light duration and speed limit, traffic congestion in a large city can be solved. Traffic congestion has been causing many challenges in most cities of modern countries. To a traveller, congestion means lost time, missed opportunities, and frustration and to an employer it means lost worker productivity, trade opportunities, delivery delays, and increased costs. By solving congestion problems it is feasible not only for physically constructing new facilities and policies but also by building information technology transportation management systems. Traffic congestion problems cannot be solved by expanding the road infrastructure. In fact, building new roads can actually compound congestion, in some cases, by inducing greater demand for vehicle travel.

K.R. Shruthi et al. [2] the proposed system efficiently utilizes and manages traffic light controllers. An adaptive traffic control system based on a new traffic infrastructure using Wireless Sensor Network (WSN). They are dynamically adaptive to traffic conditions on both single and multiple intersections. In this project an intelligent traffic light controller system with a new method for vehicle detection and dynamic traffic signal time manipulation is used. The project also controls traffic over multiple intersections and follows international standards for traffic light operations. A central monitoring station is used to monitor all the access nodes.

Yousaf Saeed et al. [3] proposed a work which presents an application of fuzzy logic for multi-agent based autonomous traffic lights control system using wireless sensors to overcome problems like speed, traffic irregularity, accidents and congestion. This agent based approach can provide a solution by minimizing the vehicle waiting time especially the emergency vehicles using fuzzy logic control under situations of emergency that normally occur. Two traffic junctions information is taken to calculate effectiveness of this system.

Mario Collotta et al. [4] in this paper a real-time knowledge of information concerning traffic light junctions represents a valid solution to congestion problems with the main aim to reduce accidents as much as possible. The Red Light Running (RLR) is a behavioural phenomenon which occurs when the driver must choose if he can cross the road or no when the traffic light changes to yellow from green. The drivers sometimes cross even during transitions from yellow to red and as a consequence, the possibility of accidents will increases. This often occurs because the drivers wait too much in the traffic and its not well balanced. In this paper we propose a technique that is based on information gathered through a wireless sensor network which dynamically processes the green times of a traffic light in an isolated intersection. The main aim is to optimize the waiting time in the queue and as a consequence reduce the RLR phenomenon occurrence.

Vikramaditya Dangi et al. [5] proposed a paper to implement an intelligent traffic controller using real time image processing. The image sequences from a camera are analyzed using various edge detection and object counting methods to obtain the most efficient technique. The number of vehicles at the intersection is evaluated and traffic is efficiently managed. This system also implements a real-time emergency vehicle detection system in which case if an emergency vehicle is detected, the lane is given priority over all the other lanes.

Jiancheng Zhang et al. [6] proposed a paper for developed a probabilistic neural network (PNN) classifier for object classification using roadside Light Detection and Ranging (LiDAR). The objective is to classify the road in urban road into one of four classes: Pedestrian, bicycle, passenger car, and truck. Five features calculated were selected to show the difference between different classes. The data were collected at three different locations that represent different scenarios. The performance of the system was evaluated by comparing the results of the PNN with those of the support vector machine (SVM) and the random forest (RF). The results showed that the PNN can provide the results of classification with the highest accuracy among the three investigated methods. The overall accuracy of the PNN was 97.6% using the testing database. The errors in the results were also diagnosed.

Feihu Zhang et al. [7] proposed a paper presenting a sensor fusion based vehicle detection approach which fuses information from both LIDAR and cameras. The proposed approach is based on two components: a hypothesis generation phase to generate positions that potential represent vehicles and a hypothesis verification phase to classify the corresponding objects. Hypothesis generation is achieved using the stereo camera while verification is achieved using the LiDAR. The main contribution being vehicle detection. This system leads to an enhanced detection performance and in addition maintains false alarm rates compared to vision based classifiers. The experimental results suggest a performance which is comparable to the current state of the art, albeit with reduced false alarm rate.

Mathias Perrollaz et al. [8] the proposed paper will detail a novel approach to compute occupancy grids from stereo-vision, and shows its application for the field of intelligent vehicles. In the proposed approach, occupancy is initially computed directly in the stereoscopic sensor's disparity space. The calculation accounts for the detection of obstacles and road pixels in the space and partial occlusions in the scene. In a second stage, this disparity-space occupancy grid is transformed into a Cartesian space occupancy grid to be used by subsequent applications. This transformation includes temporal and spatial and filtering. The proposed method is designed to be easily processed in parallel. Consequently, we chose to implement it on GPU, which allows real-time processing for the demanding application. In this paper, we present this method and we propose an application to the problem of perception in a road environment. Results are presented with real road data, comparing qualitatively this approach with others.

Madhurima Pandey et al. [9] the proposed system uses an automated intelligent system that can handle traffic easily compared to our current scenario where we provide manual power to handle it which is not possible each and every time. This paper discusses about some of the standard traffic control system and their drawback, image processing technique which helps in finding traffic queue length and some of the methods of it.

Sheena Mariam Jacob et al. [10] proposed a paper that aims to overcome traffic congestion caused by ineffective traffic management systems that are outdated and work on a predefined countdown. These traditional systems allot timings irrespective of the actual density in traffic on a specific road thereby causing large red light delays. The system we propose ensures traffic lights respond to real time values of traffic, thereby allowing proper management of time and resources. In order to do this we first calculate the density of traffic which is determined using a combination of ultrasonic sensors and image processing techniques. This information is processed by a Raspberry Pi, which in turn controls the traffic light indicators. In addition to that, the data that is collected is sent to the cloud, and can be used to monitor traffic flow at periodic intervals. In case of sensor system failure, the values stored in the cloud will also be useful in predicting the density of traffic based on long term periodic analysis.

3. LITERATURE GAPS

In [1], a new framework has been proposed for automatic traffic light control systems for improving traffic congestion problem. To automatically set the time duration of red and green light signals, a simulation model is used for improving traffic problem in peak hours according to the traffic conditions in the street. The simulation results prove the efficiency of the simulation model as the average waiting time and number of cars are dropped down sharply when the red light duration is 50 seconds. Further analysis also shows if we set the red light durations of the three intersections to 50 seconds, the total performance of the model is the best. Although this paper presents and the DATLCS, there are still several aspects where we can further improve its functions. In particular, the simulation model can be extended to add some more as well as different directions road to relate the model more close to the reality. In addition we can collect traffic flow and average car speed by using RFID technology, the method of dynamically finding a best route or a second optimal route for road navigation systems will be also a major research issue in the future.

In [2], the system can be extended by using GPS navigation system installed in vehicle and can be used to detect VIP vehicles and send message to authority.

In [3], Data is collected from the two junctions and a traffic control system for multi-agents is proposed, the emergency vehicles passes these two junctions quickly with less traffic and at the same time collisions are also avoided in case of multiple emergency vehicles coming from different directions. In case of new hardware technology and algorithms, the proposed system is flexible enough to enhance and handle future traffic aspects using FPGAs based microelectronics chips to control traffic signal lights.

In[4],the paper proposed a technique for managing traffic light cycles in order to reduce the queues in road and as a consequence, accidents due to the RLR phenomenon. Therefore, a wireless sensor network has been used with the aim to gather real time information about roads congestion. These information are later processed by a central node which is equipped with a

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 175-178, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1001

special module and is based on an algorithm that dynamically processes the traffic light cycles reducing, reduces waiting times and drivers frustration. The results, clearly demonstrate how the proposed algorithm improves the queue management around a traffic light.

In [5], The focus shall be to implement the controller using DSP as it can avoid heavy investment in industrial control computer while obtaining improved computational power and optimized system structure. The hardware implementation would enable the project to be used in real-time practical conditions. In addition, this system can identify the vehicles as they pass by, giving preference to emergency vehicles and assisting them in traffic on a large scale.

In [6], The future transportation system must rely on multiple sensors including LiDAR, camera, radar et al. The data from different types of sensors overcome the limitations of one single sensor and provide more features for classification. As an example, thermal imaging may be used for object classification considering that different users may have different temperatures. Therefore, it is also necessary to use the data from different sensors to further improve the accuracy of object classification. Integrating the data from different sensors is another research topic for future studies.

In [7], The results illustrate that the proposed approach achieves a lower false alarm rate in urban environments, which may be helpful in future autonomous navigation system. Future work concentrates on modeling contour parameters to 3D models from 2D models.

In [8] The future work are as follows. First, the algorithm must be tested extensively, in conjunction with laser scanners for Bayesian sensor fusion. Then adding sub-pixel estimation of the disparity values will provide improved accuracy, without modifying the method itself. The problem is to use an approach which provides actual separation between the road surface and the vertical objects.

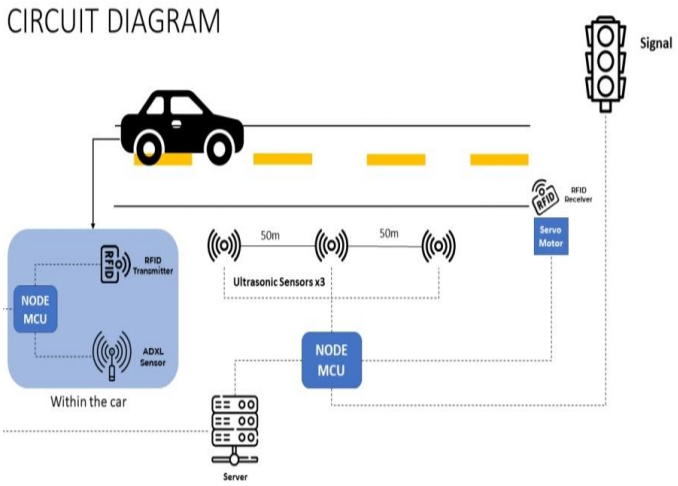
In [9] This work can be enhanced further by proposing a system which identifies the presence of emergency vehicles and by giving preference to those emergency vehicles. Secondly, it can be enhanced by using VANETs(vehicular Ad-hoe Networks) as it provides road safety and intelligent transport system.

In [10] This project still has space for improvement and can be extended by displaying traffic data in an application that can be accessed by the public. As an addition the system can be made efficient by using a camera with high resolution or by replacing the HC-SR04 ultrasonic sensors with industrial grade sensors that serve the same purpose. Further changes could also be made to the system which gives highest priority to emergency vehicles in any situation.

4. PROPOSED METHODOLOGY

This project is being designed to increase road safety even during the absence of traffic police and to bring their attention to those who break the law. The traffic is coordinated in a circular loop that takes in the inputs on real time based on the density of the vehicles on road. NODE MCU with ESP Wi-Fi module are used to transfer and collect all the data from the sensors and are made available at our local servers that are setup and will receive the data from the NODE MCU.

The proposed system utilizes the Wi-Fi capabilities and Microcontroller present in Node MCU and the accuracy achieved by Ultrasonic sensors. By pairing up 3 US Sensors at a calculated distance, we can successfully judge the densities of traffic into distinct levels. Levels can be used to control the signals vehicle movement and hence the traffic can be managed better. The addition of RFID enables to justly find all the traffic rule violators by catching them jumping a red sign and by the presence of an ADXL sensor that is paired with a Node MCU inside the car, we are able to capture data of any rash driving incidents made by the driver of the car at any point in time and on any street. Thereby, allowing the respective authorities to take further steps. The Server plays an important role of making aware all the information to the public for their general needs as well as the Local Authorities to correctly handle all the fines handed out to rule violators.



5. CONCLUSION

The proposed mechanism provides a real time traffic monitoring system that enables the respected departments to monitor traffic patterns from the cloud. Valuable data can be extracted for road planning. The live control of signals is enables such that the traffic can be maintained at all times to be at optimal levels and the altering of signal lights can be done automatically. This ensures better flow of traffic at peak hours and the best possible setting during other times based on live values. It also makes the process of finding the law breakers easier and automates the entire process where no law enforcement officer's presence would be required locally at signals or junctions. The systems provide a platform to store the data as well where all data can be accessed by the law.

6. FUTURE WORK

The system currently is developed for a single junction which takes in 2-way traffic. The system can be expanded to integrate a greater number of such junctions with higher lanes merging, enabling the system to act as a unified traffic managing hub, where data from all over the city can be collected and can be integrated for data analysis and traffic patterns. The managing of signals also becomes more streamlined when data can be used from all signals leading up to another signal, enabling the smooth flow of traffic. The system can also incorporate bigger RFID systems that can scan multiple cars braking signal at once. Higher powered ultrasonic sensors can be used for wider roads and lower powered ones for narrower roads, enabling the best suitable outcome without any power loss as such.

7. REFERENCES

1) R. M. Cardoso, N. Mastelari, and M. F. Bassora, "Internet of things architecture in the context of intelligent transportation system a case study towards a web-based application deployment," presented at 22nd International Congress of Mechanical Engineering (COBEM 2013), 2013.

2) Wen and Yang, “A dynamic and automatic traffic light control system for solving the road congestion problem” WIT Transactions on the Built Environment (Urban Transport). Vol. 89 , 2006, pp 307-316.

3) Chen and Yang, “Minimization of travel time and weighted number of stops in a traffic-light network”, European Journal of Operational Research. Vol. 144, pp565-580.

4) Pappis, C.P. and Mamdani, E.H., “A Fuzzy Logic Controller for a Traffic Junction”, IEEE Transactions on Systems, Man and Cybernetics, 1977, pp 707-717.

5) Road Traffic Congestion Monitoring and Measurement using Active RFID and GSM Technology by Koushik Mandal, Arindam Sen, Abhijnan Chakraborty and Siuli Roy, IEEE | Annual Conference on Intelligent Transportation Systems, 2011.

6) Priority Based Traffic Lights Controller Using Wireless Sensor Networks by Shruthi K R and Vinodha K, International Journal Of Electronics Signals And Systems (IJESS) ISSN: 2231- 5969, Vol-1 Iss-4, 2012

7) Image Processing Based Intelligent Traffic Controller by VikramadityaDangi, AmolParab, KshitijPawar and S.S Rathod. Undergraduate Academic Research Journal (UARJ), ISSN : 2278 – 1129, Vol-1, Iss-1, 2012.

8) Intelligent Traffic Signal Control System Using Embedded System by Dinesh Rotake and Prof.SwapniliKarmore, Innovative Systems Design And Engineering, ISSN 2222-1727 (paper) ISSN 2222-2871 (online), Vol. 3, No. 5, 2012

9) Xia and Shao, “Modelling of traffic flow and air pollution emission with application to Hong Kong Island”, Journal of Environmental Modelling & Software. Vol. 20, 2005. pp 1175- 1188.

10) Stoilova and Stoilov, “Traffic Noise and Traffic Light Control”, Journal of Transportation Research, Vol. 3, No. 6, 1998, pp 399-417.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

Machine Learning Load Balancing Techniques in Cloud Computing: A Review

Juliet Gathoni Muchori[1]
Murang'a University of Technology
School of Computing and Information Technology
Department of Information Technology
Murang'a, Kenya

Peter Maina Mwangi[2]
Murang'a University of Technology
School of Computing and Information Technology
Department of Computer Science
Murang'a, Kenya

Abstract: Load balancing (LB) is the process of distributing the workload fairly across the servers within the cloud environment or within distributed computing resources. Workload includes processor load, network traffic and storage burden. LB's main goal is to spread the computational burden across the cloud servers to ensure optimal utilization of the server resources. Cloud computing (CC) is a rapidly growing field of computing that provides computing resources as a product over the internet. This paper focuses on the issues within Cloud Load Balancing (LB) that have attracted research interest. The paper also mainly focused on uncovering machine learning models used in LB techniques. The most common algorithms in the reviewed papers included Linear Regression, Random Forest classifier (RF) Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Long-Short Term Memory- Recurrent Neural Network (LSTM -RNN). The criteria for LB technique was identified through performance metrics like throughput, response time, migration time, fault tolerance and power saving. The paper adjourns by identifying research gaps found in the reviewed literature.

Keywords: Linear Regression, Random Forest classifier (RF), Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Long-Short Term Memory- Recurrent Neural Network (LSTM -RNN), Load Balancing (LB)

1. INTRODUCTION

Load balancing refers to the distribution of the computing workload to a group of servers. Load implies CPU load, network traffic burden and server storage capacity. The workload originates from the client requests and is sent to the servers [1]. The concept of load balancing is applied in distributed system administrators to sub-divide, allocate and issue resources between different servers, networks and computers [2]. Load balancing provides the ability to cope with growing hardware architecture and computing needs. The system performance ideally remains relatively independent of the increasing input variables [2].

The importance of load balancing includes equitable distribution of computing resources by ensuring nodes are not overloaded or underloaded [1]. In return these improves the speed and overall throughput of the whole distributed system. Other critical contributions of load balancing include cloud scalability by distributing the new additional workload effectively to the new instances of virtual servers and starts different services to address the growing requests.

Other importance of load balancer in cloud environment is the detection of the idle nodes and the newly added servers. The LB component is in charge of directing new requests to the discovered servers [2]. Another importance includes the disaster recovery by having the LB component handle the cloud services continuity in case of catastrophic failures. The load balancer redirects the requests to the available servers. It exhibits transparency to the user by making all services available despite internal failures or large workloads.

A typical load balancer accepts requests from clients or other servers within a network and then assigns the most appropriate server to handle the load [3]. There are several types of load balancers [2]. They include software-defined network (SDN), User Datagram Protocol (UDP), Transmission Control Protocol (TCP), Server Load Balancer (SLB), Virtual and Load Balancer as a Service (LBaaS) [3].

Cloud computing is a rapidly growing field of computer science that is concerned with scaling distributed systems, networks and storage to mega utility computing systems [4]. Cloud service providers (CSP) provide access to both hardware and software resources to the cloud users on demand. Users access the cloud resources remotely over the internet on a rented basis. Some of the popular firms offering cloud solutions include Apple, Amazon, Google, IBM and Microsoft [4].

Cloud services can be categorized into three service models, namely: Infrastructure as a service (IaaS), Software as a Service (SaaS) and Platform as a service (PaaS). IaaS is a cloud service that offers access to Information Technology (IT) hardware and software over the internet. The host is in charge of managing and availing the infrastructure resources to the cloud user. For instance; virtual machines, servers, storage and network resources [5].

PaaS provides the resources to create, publish and customize the software in a hosted environment. The cloud user is allowed to install their software and tools like database, web server and application servers [3]. There are

specialised platforms to offer unique features like communication. For instance, a communication platform as a service offers real-time communication capabilities like video and voice. PaaS eliminates the software license costs [5].

SaaS provides the cloud user with ways of accessing software from anywhere over the internet connection. The host offers access to the application and the data for instance Yahoo!, Gmail, iCloud, Microsoft Office 365 [5]. Instead of installing the software, the user simply accesses it over the internet. It frees the users from complex handles of managing and installing the software. The cloud provider has the responsibility to manage the access, to secure and to make available the application.

Cloud computing is deployed using various methods. The design and setup of the cloud environment can follow different topologies. These topologies include; Private Cloud, Public Cloud, Hybrid Cloud and Community Cloud. Private Cloud is set up by a cloud provider and its application and infrastructure are operated by the application provider entirely. The cloud is solely dedicated to the needs of a single organization [5].

A public cloud is an open setup model where the infrastructure facilities are provided by a third party. This topology allows resource sharing between multiple organizations or consumers. The public cloud makes the computing resources available to anyone for a subscription fee. All the hardware, software and other supporting infrastructure are owned and managed by the cloud provider. This model is the least expensive choice for application hosting [6].

Hybrid cloud combines private and public cloud and it eliminates the need for the trust model. Both public and private clouds require interoperability and portability of data and applications that allow communication across the models. This model is less expensive compared to the private cloud. On-premises datacentre shares data and applications with the public cloud.

Community cloud is similar to the extranets but it has dedicated virtualization on demand. Organization sharing common goals or a specific community builds a shared cloud to be used by its members [6]. The community cloud has multiple tenants that share similar concerns in security, performance and the reach of the cloud. Services offered are limited to the computing needs and the requirements of the community members.

The host of the cloud computing services has to ensure ease of access, availability and the fastest access of the cloud services. Cloud computing is experiencing growing consumers that requires the scaling of computing infrastructure and enhancement of resilience and fault tolerance. To maintain quality cloud services, the host has to conquer load balancing, fault-tolerance, virtualization and cloud security [5].

The goal of this review paper is;
To identify major challenges facing load balancing in cloud computing.
To review machine learning-based approaches to Load Balancing.
To identify renowned metrics for load balancing.
There is a need for an intelligent burden balancer since effective load balancing solves the majority of the cloud computing performance issues. This paper focuses on the application of artificial intelligence to solve the load balancing problem.

2. RESEARCH DESIGN

This chapter describes the organization of this research paper. It goes further to discuss the set of research papers evaluated in this paper along with the sources of those papers and their elaborate search criteria.

A. RESEARCH QUESTIONS

This review paper was guided by the following main research questions:

a). What are the current major research challenges in cloud load balancing?

c). What are the various machine learning load balancing techniques that have been developed so far?

d). What are the criteria to identify an effective load balancing technique?

e). What research gaps still exist in the current load balancing approaches?

The above questions are answered by carefully considering accurate Cloud Computing and Load Balancing published peer-reviewed research papers acquired through the search criteria discussed below.

B. SEARCH CRITERIA

A step-wise analysis of the load balancing and cloud computing were conducted over well-known paper indexing engines. Some of the search strings that were used include cloud computing, load balancing, challenges in cloud computing, research challenges in load balancing, machine learning-based load balancing in CC, deep learning-based load balancing techniques in CC. The search engines used in this paper are tabulated below in table 1:

Table 1: Research Papers Search Engines

Finding Engine	Source Address
Semantic Scholar	https://www.semanticscholar.org/
Science Direct	https://www.sciencedirect.com/
IEEE Xplore	https://ieeexplore.ieee.org/
Research Gate	https://www.researchgate.net/
Google Scholar	https://scholar.google.com/

C. DATA SOURCES

During the preparation of this survey, several data sources were considered. The research paper primarily looked for conferences, periodicals and journal papers in Google Scholar, Scopus, Science Direct and other databases of related research papers.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

Table 2: Inclusion Criteria Table

Criteria	
Inclusion	• A paper that outlined the research problems present in cloud load balancing. • A research paper authored by scholars or practitioners. • A research paper that focuses on machine learning based load balancing techniques. • A peer-reviewed publication. • English written paper
Exclusion	• A research paper that doesn't emphasize intelligent load balancing techniques. • Commercial Papers that required purchase to access.

D. STRING REFINEMENT

The search strings were entered on the search engines discussed above in Table 1. Usually, the search was refined gradually after the broad search of cloud computing, then load balancing in cloud computing, machine learning-based load balancing in cloud computing and deep learning-based load balancing in cloud computing. The research papers were constrained to papers published from January 2016 and December 2021.

More than 70 papers were found with majority of them being commercial access and others lacking direct relation to our field. After eliminations around 30 papers were reviewed. The criteria were focused on the keywords. Although sometimes the criteria were further refined to include open-source papers, it reduced the relevance of the papers found and it was preferred to check the abstracts of the papers with restricted access.

E. QUALITY ASSESSMENT

On the searched research papers, quality evaluation criteria were applied for inclusion and exclusion of the research papers. An initial study of the papers executive summaries was studied and depending on our guiding research questions, the paper was included or excluded.

Thereafter, the selected research papers were fully read based on the criteria and the papers were either included or excluded for review. The inclusion criteria entailed a major focus on the load balancing techniques in cloud computing and intelligence-based techniques.The exclusion criteria were based mostly on research paper accessibility and relevance to the goals and the research questions of this review. Some papers written in the context of cloud computing could not be included since they focused on the traditional load balancing techniques. The paper will proceed to discuss the answers to the research questions by beginning with the challenges in cloud computing.

3. RESEARCH CHALLENGES WITHIN CLOUD LOAD BALANCING

The primary goal of load balancing is to ensure cloud nodes are not under or overloaded. The LB process can be described as the spread of the work burden across the network links on the multiple clusters to maximize the use of the assets and cut the overall turnaround time. Burden balancer is strategically placed within a cloud architecture. It is situated such that it receives the cloudlets requests and determines which server to forward the request to. In some cases, the load balancer is positioned as a server that performs tasks distribution to other servers [3].

An ideal load balancer should exhibit intelligent behaviour in allocating the requests smartly. Smart resource allocation involves equal load allocation on all the servers at every single point in time. The emergence and growth of artificial intelligence have shifted the load balancing research from the traditional load scheduling algorithms like min-min, round-robin to intelligent load balancing models based on machine learning and deep learning [2].

An intelligent load balancer offers a competitive advantage to the cloud providers by ensuring the quality of service and compliance to the established service level agreement (SLAs) [1]. Some of its key significance include the highest performance in terms of response and throughput, web traffic management, effective handling of the ad-hoc traffic bursts and surges, and flexibility. Load balancer offers elasticity in terms of scalable computational requests on client demand.

This section will highlight the research challenges present in load balancing, these are the problems that intelligent load balancing sought to address. Cloud management is faced with uncertainty since the resource demand keeps on charging every second. Some of the logical and physical problems that complicate load balancing include:

The physical location of the datacentres poses some logistical and response time challenges in the load balancer since cloud providers have datacentres in different continents and cloud users expect the cloud to perform without delays. Another challenge is in the edge computing problem that recommends the cloud requests be processed near where they emanate [4].

Migration of the virtual machines (VM) is another problem within load balancing since overloaded physical machines are prompted to migrate some of their virtual machines to another underutilised physical machine [7]. The challenging part is copying the current location of the VM memory pages and transferring them across the network without affecting the services its offering [8]. Live VM migration consumes a large part of the network bandwidth, CPU and memory resources that can easily violate the SLA.

The complexity of the balancing techniques raises the question of the algorithmic complexity of the load balancing technique. An efficient load balancer should not be very complex in terms of hardware requirements and fault tolerance. Good techniques make compromises to maintain optimal performance [3].

Heterogenous nodes present another research challenge to the load balancer since the nodes have different

computational capabilities, memory and networking components. The nodes have a different kinds of machine architectures (GPUs, CPUs, Multicore CPUs). Establishing a uniform mechanism to share the indifferent resources by assigning tasks poses the problem [9].

Single Point of failure occurs when the entire system depends on a single load balancer component [7]. There is a need to have a redundant load balancing component that implements a failover solution that forwards the burden-sharing component to another load balancer within the same cloud network. Some scenarios advocates for having a primary and a standby load balancer that automatically switches on failure [2].

The scalability of the load balancer is another load balancer design consideration. Cloud providers have interconnected nodes that are added to cater for the growing cloud demand. A scalable load balancer allows the addition of the hardware and scaling of the load balancing component [1]. The load balancer should maintain the performance after task scaling. It maintains a balance between the used and unused resources.

4. MACHINE LEARNING -BASED LOAD BALANCING TECHNIQUES

This section is dedicated to intelligent load balancing techniques. Load balancing techniques are categorized differently according to different features. They play a major role in server resource utilization. Load balancers are built upon some cloud environment aspects like the server CPU and memory resource, service level agreements (SLAs), prediction of the network congestion, quality of service (QoS), service response time estimation, and the storage demand within the cloud [10].

Machine learning is a part of Artificial Intelligence that focuses on training systems to perform new tasks without being explicitly programmed. Historical data and statistical techniques are combined through a process called training to build models that can forecast new unseen values [11]. Deep learning is a subset of machine learning that uses variations of neural networks with deeper networks and large datasets. Deep learning combines feature extraction and prediction in a deep network within hidden layers. It achieves better performance than traditional machine learning problems [12].

The following intelligent models were reviewed:

a. Deep learning regression technique

Deep learning-based regression was used to predict the continuous schedule of tasks from the computational time and cost by Kaur and others [13]. The deep learning network was designed to have 3 hidden layers of convolutional neural networks, a pooling layer and the activation layer made up of the ReLU function. The training data was composed of the time and cost parameters data from larger workflows.

b. Fully Connected Network (FCN) Technique

The deep learning-based load balancing mechanism was made up of fully connected convolutional layers. The model was developed by Zhu and others [14] to replace the hash functions used traditionally to schedule the tasks. Historical cluster access data was used to train the model. The FCN model was designed as a hierarchical model made up of sub-models that feed their output as input to the next hierarchy stage [14]. The hierarchy was made of 4 stages with input, disperse, mapping and join stages. Each sub-model had 3 fully connected layers [14]. The models used the deterministic approach to map the workload to the servers.

c. Support Vector Machines (SVM) and K-suggest technique

Lilhore and others proposed [15] a load balancing solution that was based on multiple machine learning algorithms like SVM and K-suggest clustering tool. Clustering is used to establish groups of virtual machines that are derived from the CPU and main memory (RAM) usage. This technique shared the assets with various groups and the VMs as well. Then it used dynamic aid mapping to assign the loads to their appropriate VM groups depending on their sizes i.e: normal, Idle, Underloaded and overloaded VMs [15]. Resource mapping involved mapping the grouped jobs with the appropriate VM group. This method improved the quality of service and lowered the overall wait or reject time [15].

d. Bayesian Network with Reinforcement Learning

Liang and others [16] proposed a load balancer to control the traffic in the Software-Defined Network Controller component of data centers [16]. The Bayesian network was used to predict the amount of load traffic and combined it with reinforcement learning for the optimal cause of action and to add a self-adjustment parameter. Software-Defined Network is the brain of the whole network that separates the data transmission layer and the control layer.

Bayesian network predicted the load traffic on the SDN controller while the predictions were used through reinforcement learning to determine best cause action [16]. Strategies adopted involved the spread of the network burden and delocalization of the processing and control. Network stability, load balancing speed and performance of the controller were achieved.

e. Regression, Random Forest and AdaBoost based technique

Machine Learning-based load distribution model [17] was made up of several models namely multiple linear regression (MLR), random forest (RF) and AdaBoost (Ada) were used to determine where each query to be processed based on the turnaround time of the CPU and GPU [17]. This technique addressed the architectural heterogeneity by accounting for the difference in the processing units and their associated performance characteristics. Its major focus was on the distributed database management systems transactions distribution [17].

f. ANN and self-adaptive differential evolution (SaDE) technique

This technique was developed by Kumar and others to predict the workload within the cloud data centre [18]. This approach combined the artificial neural network (ANN) and the self-adaptive differential evolution (SaDE). User requests were amassed to time units that were used as the historical data. The ANN part was trained with the actual workloads and the historical data [18]. The resultant model was used to forecast the upcoming work in the data centre. The model was trained on datasets from NASA and Saskatchewan servers [18].

g. Long Short-Term Memory -Recurrent Neural Network (LSTM-RNN) approach

LSTM is a special kind of RNN that preserves the weights from past activities that making it a good deep learning algorithm to be used in time-series forecasting. LSTM algorithm has an inbuilt forget gate that allows it to escape the long-term dependencies up to some point and it only keeps the necessary features. In their paper, Kumar [19] and others sort to address major issues within load balancing namely: power utilization and scaling of resources dynamically.

LSTM-RNN load balancing technique was developed by analyzing the history of the data centre through the cloudlet traffic logs with consideration to the time factor. Insights that were gained from the historical data were used to predict the future workload. The training data is continuous over a period. The projected workload data was used to expand the resources and to shut down the unused resources to preserve power [19]. LSTM-RNN model was trained on the HTTP traces of NASA, Saskatchewan Server and Calgary server datasets [19].

h. Back Propagated Artificial Neural Network (BPANN) approach

BPANN was used on a dynamic-agent based load balancer that was proposed by Prakash and Lakshmi on the software-defined network (SDN) [20]. SDN is a component within the cloud architecture that is visible globally. As part of easing the work burden within the load, they are tasked to migrate the VMs within the data centre.

BPANN algorithm was trained on the VMs load and migration data. The resultant model was used to predict the VM load. The projected load was then used to determine the VM migration. Effective VM migration improves the network efficiency and the rate of data migration. Processing speed was reduced considerably since the heavy loads are matched to the underutilized VMs [20].

i. Quantum Neural Network (QNN) approach

QNN is a variation of neural networks that are based on the principles of quantum computing. Quantum circuits have been found to behave like artificial neural networks [21]. QNN have some computational advantages since they are made up of only the needed/necessary parameters to fit specific data. That feature makes them outperform their classical counterparts [21].

QNN model was used to predict the workload to be generated by the cloudlets. Singh and others encoded the workload data in qubits and the model was used to estimate the workload and the needed resources with much precision [22]. Their model used the Controlled-NOT (CNOT) gate was used as the activation function in both the hidden and the output layers that adjusted the qubit network weights [22]. Further network weight optimization was done by a self-balanced differential algorithm [22].

Intelligent load balancing has taken over from the traditional load balancing approaches. The use of machine learning and deep learning algorithms to develop load balancing models have improved the response time, resource elasticity and conserved power. For instance, Google adopted neural networks in its data centres to manage the cooling of the centres that reduced the power used for cooling by 40% [23]. This has shown the potential of artificial intelligence in addressing complex problems.

Table 2 has summarized the algorithms used in the intelligent load distribution techniques discussed above, the kind of data used for training those models and the major problem of load balancing it has addressed. It can be noted that deep learning is taking over the traditional machine learning models since it has offered better accuracy even with the growth of datasets. Hybrid solutions are the majority of the reviewed papers, with many researchers opting to combine several algorithms to form a model for instance: [18], [22], [24] and [17].

Table 3:A summary table showing reviewed intelligent load balancing techniques

No.	Publication	Underlying Machine/Deep Learning Model	Data Used	LB Problem Addressed
1	Deep Learning Regression [13]	CNN	Tasks workflow data	Quality of Service (QoS) resource utilization and throughput
2	Deep Learning-Based Load Balancer [14]	Hierarchical sub-models of FCN	Historical cluster access logs	Solves data skew problem in classical LB
3	Lilhore machine learning-based LB [15]	SVM, K-Means Clustering	RAM & CPU usage data	VMs resource utilization & execution time reduction
4	Reinforcement based SDN controller [16]	Bayesian Network & Reinforcement Learning	Network traffic data	SDN Controller LB, Network Stability, Security
5	Distributed database query load distribution [17]	multiple linear regression (MLR), random forest (RF) and AdaBoost (Ada)	Database queries data	Cloud Heterogeneity of CPU & GPU
6	Workload prediction [18]	Artificial Neural Network and self-adaptive differential	Client request amassed to time units	Distribution of workloads

		evolution (SaDE)		
7	Temporal aware LB [19]	LSTM-RNN	Cloud workload with a time factor	Resource elasticity & power saving
8	Dynamic agent LB [20]	Backpropagation Artificial Neural Network BPANN	Network Traffic Logs	VM migration, data migration
9	Quantum based LB [22]	Evolutional Quantum Neural Network EQNN	Cloudlets workload logs	Dynamic resource scaling

5. PERFORMANCE METRICS OF LOAD BALANCING TECHNIQUES

LB performance parameters can be measured through some of its quantifiable features. The metrics can help in identifying the best approach to load balancing. Some measurable attributes are directly measured while others are dependent on related variables. The following metrics were found to be effective in rating a load distribution component:

a. Throughput

Throughput describes the measure of the number of tasks/items passing through a process in each time interval. In load, balancing throughput can be considered as the number of activities the LB component can handle in a specific period [3]. For instance, the LB component has high throughput if it responds to requests since it will handle more tasks than one with delayed response.

How the LB component forwards the request and the time it takes to decide on which cluster to assign the workload affects throughput. The number of accomplished tasks within a specified period also measures throughput. This metric was found in these papers; [17], [13], [4], [16].

b. Migration Time

The duration of transfer is the time the LB component takes to move processes from overloaded devices to underutilized devices. In load balancing, migration time is measured as the time required to move VMs from one physical machine to another. Migration is initiated when a task requires execution through multiple VMs or when a task is interrupted [3]. A higher number of migrations result in more migration time. An effective load sharing technique minimizes VM migrations. i.e. [20], [16]

c. Response Time

This metric measures the time taken by the LB module to respond to a task/cloud request. Response time is calculated by summing up the transmission time, waiting time and service time [25]. A good LB technique maintains very minimal response time since performance is inversely proportional to response time. This metric is very common; [15], [16], [18]

d. Fault Tolerance

Fault tolerance describes the ability of a system to withstand failure. In load balancing, it provides the ability to perform uninterrupted service even when some of its parts fail [3]. An effective load balancer continues to function even when some hosts, VMs and PMs fail. The ability to solve logical errors ensures fault tolerance. This performance parameter is measured by having single or multiple points of failure. Redundant LB components are advised to ensure that the LB component does not fail. For instance; [16] addresses fault tolerance by decentralizing network control.

e. Power Consumption

This metric defines the amount of energy/electricity the VMs consume after the load balancing technique has been implemented. A well designed LB approach reduces power usage in the VMs [25]. Usually, a load balancer ensures that the VM are not overloaded and hence they use less energy. More power saving oriented techniques shut down the unused physical machines/hosts [4]. Some papers that put power into consideration include; [15]and [19].

6. FINDINGS AND DISCUSSION

The focus of this paper was to find the latest trends in cloud load balancing research by uncovering the most used machine learning algorithms in load balancing components. Both traditional machine learning and deep learning models were found as an adaptation of the big data age.

Traditional machine learning algorithms found include: Multiple linear regression (MLR) and Random Forest (RF) Classifier [17]; SVM and K-Means Clustering [15]. Due to the complexity of the load balancing and the large data associated with their training like the CPU logs, network traffic data and storage logs. Traditional machine learning algorithms are being replaced by deep learning models.

Deep learning models implemented in this area include BPANN [20], CNN [13], FCN [14], ANN [18], LSTM-RNN [19]. The deep learning models has exhibited better performance in terms of predicting the accuracy. These models handled the big data very well without compromising the quality of model. These models exhibit an important trend that moves from spatial oriented models like ANN and CNN to spatial-temporal models like LSTM and CNN-LSTM. These trend shows that time as an important consideration in the load balancing process.

Other deep learning models that stood out of the rest include the deep reinforcement learning [16] based load distribution component and the Quantum Neural Network [22] based load balancer. These trends are revolutionizing load balancing by harnessing the power of reinforcement learning by incorporating self-taught agents in reward-based system that continuously improve the prediction accuracy. Quantum computing power improves the training speed and the overall model throughput.

Hybrid models have delivered better results than the single algorithm models. Ensemble and transfer learning are major types of machine learning model combination paradigms used in many papers. For instance; [16], [22], [20], [15], [14] , [19] and [18] are all hybrid models combining either several machine learning algorithms or combing machine learning with other technologies such as quantum computing [22].

The quality of load balancing can be quantified using some metrics found in the reviewed papers. Most used metrics found almost in every paper include: through put, migration time, response time, fault tolerance and power consumption. Therefore, the criteria for defining the best load balancing model should select a model that has high throughput, less migration time, least response time, ability to withstand server failures and ability to save power.

7. RESEARCH GAPS

This paper was dedicated to intelligent LB techniques within cloud computing. A lot of research has been done within this field and state of the art performance has been achieved in many papers. Some of the notable trends were how the research moved from traditional machine learning models like regression to deep learning models like ANN and LSTM [18] [19] [20]. Another important landmark is the embrace of quantum computing to solve workload prediction [22].

Most of these research works are focused on single components of load balancing. For instance, some solutions focus on network load only [16], others focus on database management query load [17] and others focus on the data centre load. There is a need for integration to have a model that links the workload problem, the VMs allocation problem, network load.

Load balancers are the single point of failure and researchers have not given fault tolerance much thought. More work should focus on this area with models having redundant distribution components. Datacenters are located within different locations; the researchers have not accounted for the network delay and incorporation of edge computing to avoid costly data transfer and VM migration over long distances.

The power conservation mechanism has not been dealt with effectively. LB techniques so far lack incorporation of energy conservation in their LB design. Techniques should put into consideration energy saving while balancing.

8. CONCLUSION

The focus of this article is cloud load balancing. Load balancing is one of the largest problems in a cloud environment. LB can adversely affect the quality of service and the SLAs hence making the cloud host lose clients. The work of the LB component is to share the work burden across the cloud resources to ensure maximum utilization of the resources and efficiency of the growing devices.

This research has identified some intelligent load balancing techniques. Models used in these papers were highlighted. For instance, multiple linear regression, AdaBoost, random forest, k-suggest, k-means clustering, CNN, FCN LSTM, ANN, QNN and reinforcement learning. These models have shown the changing trend from traditional machine learning [17], to deep learning [19] components embedded with reinforcement learning [16] to foster continuous learning and finally harnessing of quantum computing power for cloud management [22].

REFERENCES

[1] DSM, "A Beginner’s Guide to Cloud Scalability and Load Balancing," DSM, 16 NOV 2018. [Online]. Available: https://www.dsm.net/it-solutions-blog/a-beginners-guide-to-cloud-scalability-and-load-balancing. [Accessed 14 Feb 2022].

[2] A10, "What is a Load Balancer and How Does Load Balancing Work?," A10, 2022. [Online]. Available: https://www.a10networks.com/glossary/what-is-a-load-balancer-and-how-does-load-balancing-work/#:~:text=Even%20though%20a%20load%20balancer%20solves%20the%20webserver,both%20front-ending%20the%20same%20group%20of%20web%20servers.. [Accessed 14 Feb 2022].

[3] M. A. Shahid, N. Islam, M. M. Alam, M. M. Su’ud and S. Musa, "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach," IEEE Access, vol. 8, no. I, pp. 130500 - 130526, 2020.

[4] T. Khana, W. Tian and R. Buyya, "Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions," CoRR, vol. arXiv:2105.05079v1, no. 1, 2021.

[5] Data Flair, "Cloud Computing Tutorial for Beginners – Learn Cloud Computing," Data Flair, 2020. [Online]. Available: https://data-flair.training/blogs/cloud-computing-tutorial/. [Accessed 17 Feb 2022].

[6] S. Parida and B. Panchal, "Environment, An Efficient Dynamic Load Balancing Algorithm Using Machine Learning Technique in Cloud," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 4, no. 4, pp. 1184-1186, 2018.

[7] A. A. Alkhatib, A. Alsabbagh, R. Maraqa and S. Alzubi, "Load Balancing Techniques in Cloud Computing: Extensive Review," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 2, pp. 860-870, 2021.

[8] M. Liaqata, S. Ninoriya, J. Shuja, J. Shuja and A. Gani, "Virtual Machine Migration Enabled Cloud Resource Management: A

Challenging Task," *CoRR*, vol. abs/1601.03854, 2016.

[9] B. Sahoo, S. K. Jena and S. Mahapatra, "Load Balancing in Heterogeneous Distributed Computing Systems using Approximation Algorithm," in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, Athens, 2013.

[10] D. A. Shafiq, N. Jhanjhi and A. Abdullah, "Load balancing techniques in cloud computing environment: A review," Journal of King Saud University – Computer and Information Sciences, vol. 02, no. 007, 2021.

[11] IBM Cloud Education, "What is Machine Learning?," IBM Cloud, 15 July 2020. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Accessed 17 February 2022].

[12] IBM Cloud Education, "What is deep learning?," IBM Cloud, 1 May 2020. [Online]. Available: https://www.ibm.com/cloud/learn/deep-learning. [Accessed 17 February 2022].

[13] A. Kaur, B. Kaur, P. Singh, M. S. Devgan and H. K. Toor, "Load Balancing Optimization Based on Deep Learning Approach in Cloud Environment," *I.J. Information Technology and Computer Science*, vol. 3, no. I, pp. 8-18, 2020.

[14] X. Zhu, Q. Zhang, T. Cheng, L. Liu, WeiZhou and J. He, "DLB: Deep Learning Based Load Balancing," *CoRR*, vol. 1910, no. 08494V4, 2021.

[15] U. K. Lilhore, S. Simaiya, K. Guleria and D. Prasad, "An Efficient Load Balancing Method by Using Machine Learning-Based VM Distribution and Dynamic Resource Mapping," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 7, pp. 2545-2551, 2020.

[16] S. Liang, W. Jiang, F. Zhao and F. Zhao, "Load Balancing Algorithm of Controller Based on SDN Architecture Under Machine Learning," *Journal of Systems Science and Information*, vol. 8, no. 6, pp. 578-588, 2021.

[17] A. Abdennebi, A. Elakas, F. Taşyaran, E. Öztürk, K. Kaya and S. Yıldırım, "Machine learning-based load distribution and balancing in heterogeneous database management systems," *Concurrency and Computation*, vol. 34, no. 4, 2021.

[18] J. Kumar and A. K. Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution," *Future Generation Computer Systems*, vol. 81, no. C, pp. 41-52, 2019.

[19] J. Kumar, R. Goomer and A. K. Singh, "Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters," *Procedia Computer Science*, vol. 125, pp. 676-682, 2018.

[20] S. WilsonPrakash and P. Deepalakshmi, "Artificial Neural Network Based Load Balancing On Software Defined Networking," in *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Tamilnadu, India, 2019.

[21] A. Abbas, D. Sutter and S. Wörner, "The power of quantum neural networks," IBM, 2 July 2021. [Online]. Available: https://research.ibm.com/blog/quantum-neural-network-power. [Accessed 16 Feb 2022].

[22] A. K. Singh, D. Saxena, J. Kumar and V. Gupta, "A Quantum Approach Towards the Adaptive Prediction of Cloud Workloads," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, pp. 2893-2905, 2021.

[23] W. Knight, "Google just gave control over data center cooling to an AI," MIT Technology Review, 7 August 2018. [Online]. Available: https://www.technologyreview.com/2018/08/17/140987/google-just-gave-control-over-data-center-cooling-to-an-ai/. [Accessed 17 February 2022].

[24] X. Sui, D. Liu, L. Li, H. Wang and H. Yang, "Virtual machine scheduling strategy based on machine learning algorithms for load balancing," *EURASIP Journal on Wireless Communications and Networking volume*, vol. 160, no. 1, 2019.

[25] S. K. Mishra, B. Sahoo and P. P. Parida, "Load balancing in cloud computing: A big picture," *Journal of King Saud University – Computer and Information Sciences*, vol. 32, pp. 149-158, 2020.

[26] GeeksforGeeks, "Top 10 Cloud Computing Research Topics in 2020," GeeksforGeeks.com, 26 September 2020. [Online]. Available: https://www.geeksforgeeks.org/top-10-cloud-computing-research-topics-in-2020/. [Accessed 17 February 2022]

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 187-192, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1003

A Survey of Awareness of Social Engineering Attacks to Information Security Management Systems: The Case of Kibabii University Kenya

Samwel Mungai Mbuguah
Kibabii University
Kenya

Tobias Okumu Otibine
Kibabii University
Kenya

Abstract: Computer based systems are socio-technical systems in nature. The security of the system depends both on technical aspect and also social aspect. The social aspect refers to people in contact with system commonly referred to as wetware. To attack the system you may consider to target the technical or wetware. Social engineering is based on exploiting human traits that make human susceptible to these attacks. The aim of this paper was establish how aware the staff of Kibabii University were of these attributes and how these attributes could be used by social engineers to penetrate the Information Security Management systems at the institution. A survey research was adopted with a questionnaire being developed using Google application, and was administered online to all staff members of Kibabii University. A descriptive analysis was carried out on feedback. The finding was that to a large extent the sampled staff are aware of these traits but there is need for awareness training to enhance the information security management system of Kibabii University

Keywords: Social engineering, attack, wetware, human traits. Information security management system

1. INTRODUCTION

The increased dependency on reliable data communication networks has created a need for ever increasing computer security. Many technological options exist for security in both hardware and software and these implementations pose formidable threats for hackers. However social engineering bypasses the electronic security measures and targets the weakest component of networks - the human users [1].

Susceptibility to social engineering attacks stems from a lack of formal security management as well as limited education regarding social engineering. Computer security organizations are pushing for increased defenses against social engineering (Allen 2004), but until the general business community realizes the threat, very little will be done to implement policies to protect themselves compared to the efforts made to establish electronic safeguards against traditional hacking techniques. Kvedar et al. [1] carried out some research with the aim of proving the viability of social engineering as a method of network attack, as well as display the need to increase education and implement measures to protect against such an attack.

Computers are designed to provide an unconditional response to a valid instruction set. The same instruction set is used to create different layers of security privileges for different category of users. Social engineering supersedes the explicit nature of machines and focuses on human emotion and tendency. Wetware has been coined to represent the human attached to the computer. Wetware is just as vital to the computer's security as any hardware or software [2]. It is this wetware that social engineering exploits.

Computers can completely secure information to prevent unauthorized access. This could easily defeat the goal of having information from being readily accessible when needed by privileged users. The goal for a social engineer is to manipulate these authorized users to gain access to privileged information. Dolan considers social engineering as the "management of human beings in accordance with their place and function in society"[3].

Social engineers prey on humans' desire to be helpful, tendency to trust people, fear of getting in trouble, and willingness to cut corners. They have found out that exploiting weakness in human nature is much easier than exploiting flaws in encrypted software. Instead of physically breaking into bank's safe, it is much easier if one can get the lock pin combination code from a bank worker (Mbuguah & Wabwoba 2015)[4].

Allen avers that the four phases of social engineering are: information gathering, relationship development, execution, and exploitation [2]. During the first phase, information gathering, information about a company is gathered with the aim of finding weaknesses that can be exploited and ways of avoiding arrest within the organization. The second phase, relationship development, rapport and trust are developed with the contact person within the organization. The third phase is actual execution of the attack where the information is actually exchanged. Finally, the last phase is utilizing information.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 187-192, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1003

Thornburgh [5] says that an attack is successful only if the target feels compelled to give up the information in spite of their gut instinct. While Manske [6] says that a successful attack bypasses anything that would be in place to ensure security, including firewalls, secure routers, email, and security guards. This causes unrest and beats the security of encryption.

Winkler and Dealy[7] provide advice on how to secure a network against social engineering. The list includes not relying on common internal identifiers within an organization, implementing a call back procedure when disclosing protected information, implementing a security awareness program, identifying direct computer support analysts, creating a security alert system, and social engineering to test an organization's security. Dolan [3] beef up the list by adding; password policies, vulnerability assessments, data classification, acceptable user policy, background checks, termination processes, incident response, physical security, and security awareness training.

Social engineering tactics include impersonation of an important user, third-party authorization, in person attacks, dumpster diving, and shoulder surfing. Dumpster diving involves sifting through a target's waste in search of critical information. However shredders should be used to shred any documents destined to the dustbin. Shoulder surfing is a basic social engineering attack based on attempts to steal passwords and login information by watching a user input the data. This especially true in automated teller machine (ATM) halls, where users do not take precaution to block any other users from seeing them keying their pin numbers. The result is that a lot of clients have lost their funds. One person lost some money from his MPESA account when he unknowingly let a young man know his pin number. The young man, picked the phone and transferred money from the person's account to his. However forensic audit helped track down the culprit [4].

Attackers prefer to remain unidentifiable to protect themselves, some tell-tale signs of an individual attempting a social engineering attack include refusal to give contact information, rushing the process, name-dropping, intimidation, small mistakes, and requesting forbidden information or accesses.

Reverse social engineering tact involves creating a situation where the targeted individual actually seeks the attacker for assistance, which provides the attacker with the opportunity to establish trust [7]. A common tendency in human nature is for one to feel indebted to their benefactors. Reverse social engineering preys on this tendency. Not only does the target trust the individual, but also feels indebted to the attacker, and will share out information he may not otherwise share out to settle that debt.

In Kenya people have been conned by people pretending to be business men expecting a certain a transaction to go through [3]. After they have developed rapport with the victim they initially ask some money before gradually increasing the amount then finally logging off, leaving the victim high and dry. Another type of fraud executed by Kamiti maximum prisoners in Kenya is to exploit the greed of their victim. They call the victim informing them that they have won some lottery. They require some information from them, including their MPESA pin numbers. Only for the victim to realize that the conmen have cleared what money they had in their accounts. Once again audit trail by service provider Safaricom Ltd[8] located the location of the scam to Kamiti and other prisons in Kenya

2. RELATED STUDIES

One of key study was entitled Understanding Scam Victims: Seven Principles For Systems Security. The researchers tried to find out on the psychology of scam victims Al, L. E. (2009[9]). Researchers then identified traits that make people vulnerable to scams. These traits were published in ACM vol 54 journal as shown in table 1.

Table 1:Understanding Scam Victims: The Seven Principles

Principle	Cialdini (1985-2009)	Lea et al, (2009)	Stajano-wilson (2009)
Distraction		~	X
Social compliance(Authority)	X	-	-
Herd (Social proof)	X		-
Dishonesty			X
Kindness	~		X
Need and greed (Visceral Triggers)	~	X	-
Scarcity (related Time)	X	-	~
Commitment and Consistency	X	-	
Reciprocation	X		~

~ -------Lists a related Principle

Also lists this principle
X First identified this principle

Wilson [10] says that the finding's support their thesis that systems involving people can be made secure only if designers understand and acknowledge the inherent vulnerabilities of the human factor. Their three main contributions were: First hand data not otherwise available in

literature; Second they abstracted seven principles; Third they applied the concept to more a general system point of view.

They argued that behavioral patterns are not just opportunities for small scale hustlers but also of the human component of any complex system. They suggested that system security architect should acknowledge the existence of these vulnerabilities as unavoidable consequence of human nature and actively build safeguards to prevent their exploitation Wilson, [10] However they did not attempt to model the relationship between the traits and system attackability [11].

The identified human traits are dishonesty, social compliance, Kindness, Time pressure, Herd mentality, greed/need and distraction. Personality traits models do exist. Researchers have identified traits that make human beings susceptible to social engineering attacks and have extended this to system view. Researchers have also identified that the human being is the weakest link in system security [11]

Mbuguah et al.[11] did extend these concepts by not only modeling the traits as applied to software systems but also introduced some metrics that are theoretically and empirically sound. He also published algorithm for determination of these metrics.

Cyber criminals have extremely targeted eCommerce as they receive and use money, relay in technology, outsourced services and use of payment technologies like mobile money and online banking channels to carry out their day-to-day transactions. Criminals have shifted to use of social engineering as it easy to exploit user's natural inclination as compared to hacking[12].

Ntubini[13] study led to the development of the Mobile Money Social Engineering (MMSE) detection framework that aids mobile users in detecting against social engineering threats that occur via Voice Calls and SMS.

Safaricom in their 2021 report[14] highlight how they have been supporting their customers to tackle fraud Identity theft and social engineering fraud have been some of the most common forms of fraud targeted at our M-PESA customers. In FY21, they continued with their customer fraud awareness drive. They highlighted the issues through an above-the-line campaign under the tag Jichanue and Take Control, using radio, TV and digital channels. With the aim to reach all customers, we sent out over 63 million SMS broadcasts. Additionally, our digital channels reached 9.5 million people/

From the related study there is to assess the level of awareness of social engineering attacks at Kibabii University.

3. METHODOLOGY

For this paper a survey methodology consisting of twenty questions was administered online to Kibabii University staff through their email addresses. The number of staff members are three hundred and thirty (330) and respondents were thirty three (33) which constituted about 10% which is an appropriate sample size [15). The questionnaire was set on Google application. Questions were set out and the participant requested to respond by clicking on appropriate button. On completion participant pressed a submit button to relay the information back to the researchers. The application did compute the percentages for each response. Test retest was applied to seven attributes and average score computed. Hence descriptive analysis was done whose findings are represented section 4.

4. RESULTS AND DISCUSSION

In this section we highlight the results of study, interpretation of the results and finally a discussion.

4.1 General information

a) Question one was on the gender composition of the respondents. The results were that of sample population 63.6 % were male while 36.4 percentage were females as shown in Fig.1

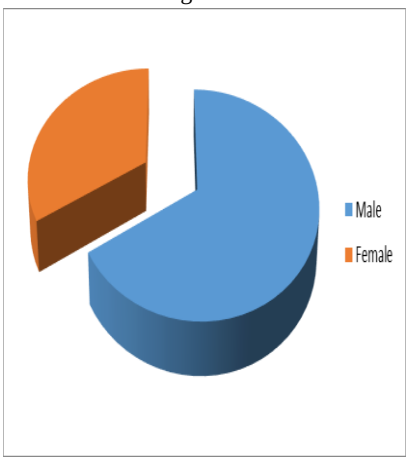


Figure 1:Gender

b) Job Category
The distribution of the respondents as far job category was:
Administrative – 48.4%, Technical – 30.3% and Academic – 21.2 %

c) The question sought to find out whether the staff knew who a social engineer was and only 60.6% could correct define a social engineer while 39.4 % could not.

d) Whether people seek the identification of strangers before serving them by requesting for ID or gate pass. 87.5% did while 12.5% did not.

e) This Question sought to find out whether they could allow a visitor mess up in their office whether the

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 187-192, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1003

visitor had some identification document or not. 97% declared they could while 3% could take no action.

4.2 Seven Attributes.

a) Social compliance-a tendency for people to obey authority or do as required of them by their superior or people in authority. The question was to find out whether the members of staff were aware that this trait could be exploited by conmen to take advantage of them. Table 2 shows the results.

Table 2:Social Compliance

QUESTION	Strongly Agrees	Agree	Do not Know	disagree	Strongly
7	90.9	9.1	0	0	0
14	24.2	24.2	12.2	24.2	15.2
9	24.2	18.2	21.2	27.3	9.1

For this attributes the positives that strongly agrees and agree (100+ 48.4+ 42.4 = 190.8)

The average 190.8/3 = 63.6

The result indicates that 63.6 % are aware that social compliance can be exploited by con artist to penetrate systems. 36.4 % are not aware. This is higher percentage that can be easily exploited; hence the need of training to enhance the awareness.

b) Time pressure-a trait of a psychological urgency attributed to insufficient time for completing required tasks. The question wanted to find out whether the participants were aware that conmen could take advantage of them by hurrying them.Table 3 shows the result

Table 3:Time Pressure

QUESTION	Strongly Agrees	Agree	Do not Know	disagree	Strongly
8	78.8	21.2	0	0	0
13	30.3	51.5	6.1	9.1	3
15	42.2	33.3	6.1	9.1	9.1

This gives a total of 257.3 and an average of 85.8%.

This means that 85.8% of the staff members are of the effect of time pressure but 14.2% are not aware. There is need for training to reduce this gap.

c) Kindness- compassion. The trait of a person having a high level of agreeableness in a personality test, usually the person is warm, friendly, and tactful. Or having an optimistic view of human nature and getting along well with others. The trait could be used by conmen to take advantage of them. Table 4 shows the result of the responses

Table 4:Kindness

QUESTION	Strongly Agrees	Agree	Do not Know	disagree	Strongly
11	81.8	15.2	0	3	0
16	27.3	42.4	6.1	18.2	6.1

The average for the positive or correct answer 83.3% and 16.7 % are not aware. There is need for training to breach this gap.

d) Greed/Need-Greed refers to a human trait of wanting more and more of something. While need is the want of something urgently and desperately. This trait can never be exploited by conmen breaking into information security systems. Table 5 shows the result.

Table 5:Greed/Need

QUESTION	Strongly Agrees	Agree	Do not Know	disagree	Strongly
12	63.6	33.3	3.1	0	0
17	42.4	30.3	0	9.1	18.2

The participant who responded positively were 84.8% and negatively 15.2%. There is need for awareness training.

e) Herd Mentality-the trait of a tendency for an individual to follow group thinking. To do something because most people are doing the same even though this may be against their better judgment. This trait could be negatively exploited by conmen to take advantage them. Table 6 show the results.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 187-192, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1003

Table 6:Herd Mentality

QUESTION	Strongly Agrees	Agree	Do not Know	disagree	Strongly
10	21.2	48.5	12.1	15.2	3
18	51.5	33.3	3	12.1	0

The Positives responses were 77.25% and negative 22.75%. The aspect of herd mentality requires more training.

f) Distraction. The trait when a secondary task obstructs/slows the user from efficiently and effectively fulfilling the time-critical main task. This trait could be negatively exploited by conmen to take advantage of them. Table 7 is representation of the results

Table 7: Distraction

QUESTION	Strongly Agrees	Agree	Do not Know	disagree	Strongly
6	90.9	9.1	0	0	0
19	36.4	51.5	3.0	3.0	6.1

The positive were at 81.85% and negative were at 18.15%. There is need for training to reduce this gap.

g) Dishonesty – the trait of being not truthful or cheating. This trait could be negatively exploited by conmen to take advantage of them in penetrating security barriers. Table 8 and figure depict the results

Table 8: Dishonesty

QUESTION	Strongly Agrees	Agree	Do not Know	disagree	Strongly disagree
20	66.7	30.3	0	3	0

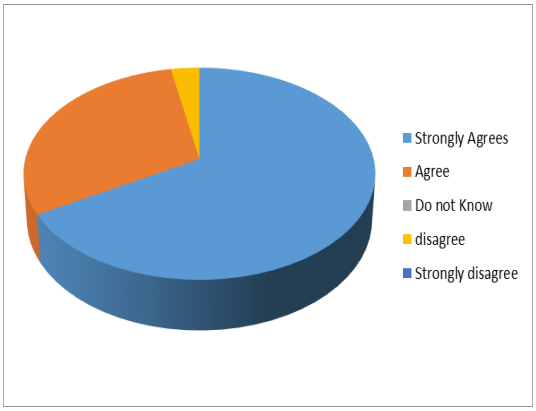


Figure 2

People appear to appreciate that dishonesty can lead to social engineering attack. The positive respondent was at 97% while the negative was at 2%.

4.3 Discussion

From the finding it evident that the staff appreciate the issues that can allow social engineer gain access to system and execute a social engineering attack. The highest concurrence being that being dishonest could easily lead to social engineering attack. It important that Kibabii maintains and up scales to 100% rating on all aspects of the human traits. However, social compliance had a concurrence of 63.6 %. It important that though it's important to obey authority from finding, this a route that the social engineer can use. There is need to continue training staff on this and other aspects. Other aspects had small but significant number not aware that a given trait could be used by the social engineer to penetrate the system. There is therefore need for continuous training and enhancement of the information security management system. If possible then certification on this standard could be an added advantage because they will lead to continuous internal and external audit of the information security management system

5. CONCLUSION

We can conclude that in general the sampled staff are to a large extent aware of the human traits that can make one susceptible to social engineering attack. However there is still a significant mass that requires further awareness training to reduce the vulnerabilities of the Kibabii University system. Everybody should be fully aware of the ever changing scenario of attacks to make the system impenetrable.

The recommendation is further training for members of staff plus further monitoring of systems including penetration testing enhancement of information security management system.

6. ACKNOWLEDGEMENT

I wish to recognize members the members of staff who took time to participate in survey. The Kibabii international conference where the paper was initially presented as conference paper.

7. REFERENCES

[1]1 Kvedar D., Nettis M & Fulton S.P(2010). The Use of formal Engineering techniques to identity weaknesses during computer Vulnerability competition .*United,States Air force Academy*

[2] Allen, M. 2004. Social Engineering: A means to violate a computer system from ,http://securitytechnet.com/resource/security/hacking/1365.pdf

[3] Dolan, A. 2004. Social engineering. SANS Reading Room. Retrieved November , 2011from http://securitytechnet.com/resource/security/hacking/1365.pdf.

[4] Mbuguah S.M. & Wabwoba F.(2015) Attackability Metrics Model for Secure Service oriented ,Architecture published by Lambert ISBN 978-3-659-66885-2

[5] Thornburgh T.(2004). Social Engineering: the Dark Art: *In the proceddingof the 1st conference on information curriculum development, GA, 133-135*

[6] Manske K.(2000). Amn Introduction to Social Engineering. Information Security Journal: A Global perspective 9:1-7

[7] Winkler I & Dealy B.(1995) Information Security Technology. Don't rely on it . A case, Study in Social engineering. *In Proceeding of the 5th USENIX/UNIX. Symposium, Salt Lake City Uta*

[8] Safaricom Ltd https://www.google.com/search?client=firefox-b-d&q=safaricom+ltd

[9] Al, L. E. (2009) Al, L. E. (2009). *The Psychology of Scams:Provoking and Commiting Errors Of , Judgement.* London: University of Exeter School

[10] Wilson, F. S. (2011) Wilson, F. S. (2011). . Understanding Scam Victims:Seven Principles For ,system ,Security. *Communication Of ACM, Vol 54,No3*

[11] Mbuguah S.M. Mwangi, W. Song P.C, Muketha G.M. (2013) Social attackability metrics in the ,*International journal of information technology research ISSN-2223-4985* Volume 3 , No. 6

[12] Nturibi LM (2018) Titled Mobile Money Social Engineering framework For Detecting Voice & , Sms Phishing Attacks - A Case Study Of M-Pesa Masters Project Report United States . International University –Africa

[13] Mwasambo LM(2016) Social Engineering In E-Commerce Platforms In Kenya MSC Project Report , University of Nairobi

[14] Safaricom 2021 : suistanable Business report https://www.safaricom.co.ke/images/Downloads/Safaricom_2021_Sustainable_Business_Report.pdf

[15] Mugenda, O. M. and Mugenda A.G (2003). *Research Methods.* Nairobi: ACTS.

[

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

Refining Location-Aided Routing (LAR) through Proactive Algorithm

Mutuma Ichaba
KCA University
P. O. Box 56808 – 00200
Nairobi, Kenya

Felix Musau
Riara University
P.O. Box 49940 – 00100
Nairobi, Kenya

Abstract: One of the weaknesses in Location-Aided Routing (LAR) is the delay due to partial flooding of data packets throughout the ad hoc network during route discovery. Systematic literature review indicates that very little or no studies conducted to seek a solution to this routing weakness in LAR. This study proposes introduction of periodic updates of location information among the nodes as a solution to minimizing latency. Proactive-LAR (P-LAR) eliminates partial flooding, thus reducing latency while advancing routing performance of traditional LAR. As a research scope, this study uses Angle of Arrival (AoA), Time of Arrival (ToA), Time Difference of Arrival (TDoA) and the expected distance of nodes and the direction of movement as the only location information details. Moreover, the simulation on OMNET ++ is limited to the initial expected zone of LAR Scheme 1. Simulation of the modified LAR Scheme 1 algorithm indicates that inclusion of proactivity as an algorithmic aspect of LAR augments general data packets throughput, latency, packets delivery ratio while minimizing the number of packets dropped. Nodes mobility in simulation was considered stationary. The simulation results—as analyzed through RapidMiner, suggest that proactive algorithmic element in LAR routing algorithm can potentially minimize partial flooding thus improving routing performance while minimizing routing overheads such as jitter and delay.

Key Terms; Mobile Ad Hoc Networks (MANETs), Flooding, Proactivity, Latency, Location -Aided Routing (LAR)

1. Introduction

A Mobile Ad Hoc Network (MANETs) is a type of wireless network that relies on IEEE Wifi standards 802.11a/b/g/n/ac/ax. A key feature of MANETs is their flexibility and the ability of the network nodes/devices to either join or leave the network at will [1]. They are composed of wireless host devices whose transmission is broadcast to any other device or node within its range. Devices out of range are reachable through hopping from along the nearest nodes. However, such flexibility comes with its challenges. Studies have established that autonomous mobility within a MANET introduce security weaknesses. Because MANETs operate without a centralized operation, it is very hard to guarantee security relative to other networks such as ethernet or fiber [2].

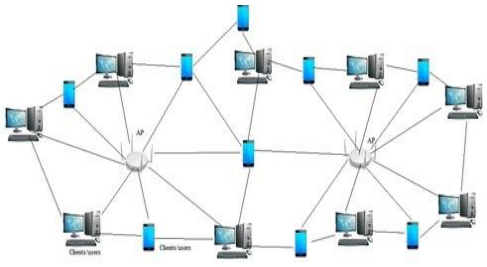


Figure 1: Example of Mobile Ad Hoc Network (MANET)

In figure 1, for example, the radio signal relay is characterized by multihopping as it is propagated from one neighboring node to another. Multihopping, however, introduces regular and unplanned network breakages and shutdowns. High mobility in MANETs results into constrain on the available resources such power and bandwidth. Besides these shortcomings, MANETs, unlike the infrastructure dependent networks, can be deployed instantly to every situation that may arise. Such networks are used to create connectivity responses for the "purpose" of a particular scenario [3]. Scenarios differ, and so do the required MANETs. For instance, a MANET formed to respond to a rescue mission is different from a MANET created for a military operation.

Other Applications of MANETs include community wireless, distributed and collaborative computing, mesh networks and multi-hop cellular networks. Constant devices movements within a MANET introduces high level dynamism the topology of the network. Although MANETs are flexible and dynamic, they experience certain shortcomings [4]. For instance, relative to other networks, MANETs experience high data packets drop rates, traffic collisions and higher signal-to-noise ratio. Other MANETs shortcomings include hidden and exposed terminals and difficulty in their modelling approaches.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

Routing protocols in MANETs are designed to respond to the network flexibility and topological dynamism introduced by the autonomous nature of network nodes. Because MANETs allow high nodes mobility and autonomy to either join or leave a network, the designed routing protocols are designed to be flexible and responsive enough to support such network dynamism. Consequently, MANET routing protocols are classified based on various criteria. Within the routing protocols, there are algorithmic rules that set out how a data packet should be moved from the source node to the destination. Routing parameters compose a key feature in MANETs routing [5].

Each of the routing protocols in MANETs has strengths and weaknesses. Among the most studied routing protocols is Location-Aided Routing (LAR). While it outperforms various proactive, reactive and hybrid routing, it experiences delay and higher consumption of power due to partial flooding. To solve this problem, this study proses introduction of proactive algorithmic element into conventional LAR Scheme 1 to combat partial flooding. The hypothesis is that proactivity will minimize the need for partial flooding of data packets during route discovery and data transmission process.

2. Routing in MANETs

Nowadays, there are a myriad of MANETs routing protocols designed for maximization of resources utilization. Various MANETs routing protocols are proposed to minimize overheads while maintaining optimal performance—data packets transmission among nodes. For instance, in figure 2, some routing protocols are designed to minimize power consumption, bandwidth, handle link failure and respond to the overall dynamism within the network [6]. Routing in MANETs are classified based on various strategies. Because MANETs do not depend on any fixed infrastructure, every mobile device/node operates as both a transmitter and a router. For this reason, each node in a MANET can receive, store, update, maintain and transmit data packets. Generally, routing is one of the most challenging tasks in a MANET [7].

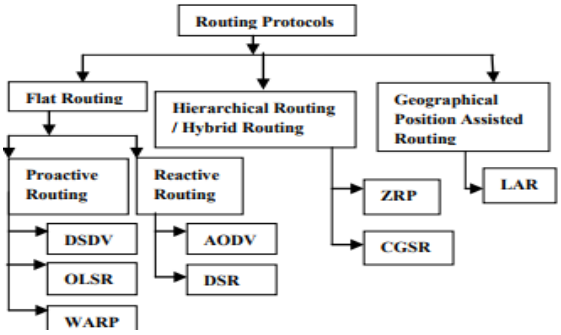


Figure 2: Routing Classification in MANETs (7)

Network engineers and administrators of MANETs have to contend with the management of limited resources while designing routing in MANETs. Network specialists have to put into consideration all the downfalls of MANETs in regard to power utilization and unplanned link failures [8]. In MANETs, the process of route discovery and their associated maintenance is paramount in network designs and developments. Routing in MANETs should exhibit certain characteristics [9]. For example, an effective routing protocol should be able to avoid signal loops while maintaining an acceptable level of quality in signal. Moreover, routing in MANETs should be able to optimize utilization of the available power, bandwidth and memory. Because MANETs are created for response in specific cases or scenarios, MANETs routing adaptability is critical. Adaptability in MANETs requires a routing protocol to be able to fully and equally distribute the network signal among the nodes [10].

Reliability of connectivity is also a trait that should be reflected in MANETs routing. All these qualities ensure reliability and stability of the network. In general, MANETs' routing protocols possess and exhibit abilities such as tolerance to faults and failures, implement ability and simplicity [11]. In addition, routing in MANETs should be able to exhibit reliability and scalability. Routing protocols in MANETs are expected to be dynamic enough to offer sufficient distribution signal in a format that allows effective and easy implementation and maintenance. Classification of routing protocols in MANETs is important. Mainly, MANETs routing protocols can be grouped based on their responses to route discovery requests—routing strategy, or based on topological/structural description of the network. Grouping of MANETs routing protocols based on routing strategy yields proactive and reactive classes [12], [13], [14].

Proactive routing protocols depend on network topological routing tables to initiate route discovery processes. Oppositely, reactive protocols rely on the source nodes to initiate route discovery requests. In proactive routing, there is constant updates of routing tables. For this reason, one of the key shortcomings in proactive routing protocols is the increased overhead in the form of power [15]. Constant updates of routing tables make proactive protocols consume more transmission power in comparison to reactive routing protocol. However, proactive routing protocols can transmit data packets faster than the reactive routing protocols. Faster data packets transmission can increase the general data throughput [16]. The trade off, however, is the higher consumption of power. Reactive routing protocols initiate route discovery process only when requested to do so [17].

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

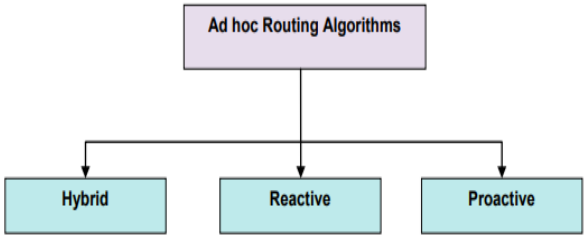


Figure 3: Classification of Routing in MANETs based on Topological Structure [18]

Consequently, the dormant nature reactive protocols conserve power but may experience slower data transmission rates. Proactive routing protocols include Destination Sequenced Distance Vector (DSDV) protocol, Distance vector (DV) protocol and Fisheye State Routing (FSR) protocol. Reactive routing includes Dynamic Source Routing (DSR) and *Ad-hoc* On-demand Distance Vector (AODV) routing [18]. When routing features of proactive and reactive routing are combined, a hybrid routing is produced. An example of a hybrid routing is the Zone Routing Protocol (ZRP). In ZRP the interior zones use proactive routing while the exterior zones as determined by the zonal radius use reactive routing.

MANETs routing can also be classified as either geographical, multicast, power-aware, hierarchical or flat based on the topological/structural makeup of a network. Most of the routing protocols in MANETs fall under the flat routing because all proactive and reactive routing protocols are included. Hierarchical routing is made up of hybrid Cluster Switch Gateway Routing (CGSR) and Zone Routing Protocol (ZRP). Geographical routing is composed of Location-Aided Routing (LAR) and Greedy Perimeter Stateless Routing (GPSR)—refer to figures 2 and 3.

3. Related Work

There is plenty of studies attempting to improve the current MANETs routing protocols. Many of the available studies examine various methods and approaches to minimize routing overheads. Comparatively, however, existing studies on LAR mainly concentrate on performance analysis relative to other routing protocols. Consequently, there is very little or no improvement proposal studies conducted on LAR. For example, study [19] attempts to compare overall performance between LAR and the Ad-hoc on-demand distance vector (AODV).

Some of the parameters used in comparative studies include throughput, delay, jitter and drop rates. Simulations in such studies are carried out under varied number of nodes and area. Study [19] analyzes the performance of LAR in as a routing protocol for a Vehicle Ad Hoc Networks (VANETs). The performance analysis in the study uses LAR to simulate vehicular movements in a city. While [20] attempts to propose a new version of LAR--Energy Efficient Location Aided Routing (EELAR) Protocol, it is not clear how the proposed algorithm saves energy. The study compares the performance of EELAR with AODV, LAR, and DSR.

Additionally, the measure of control packet overhead as an overhead is not clear enough to allow such comparison. According to [20], cluster routing can improve routing LAR. The Cluster Based Location-Aided Routing Protocol for MANET (C-LAR), outperforms the traditional LAR because it introduces scalability and effectiveness. According to the simulation results, C-LAR outperforms traditional LAR on routing overhead, delay and packets collision.

However, the paper does not show how the modified algorithm alters the technical performance of the traditional LAR in different circumstance. Other studies concentrate on general performance of LAR without algorithmic modification or performance comparisons. For instance, studies, [21], [22] analyze routing general performance of LAR. Major differences in generalized LAR routing performance studies include variations in simulators, parameters and situations.

4. Location-Aided Routing (LAR)

Proposed by [23], Location-Aided Routing (LAR) seeks to improve reactive routing protocols such the Dynamic Source Routing (DSR) and Ad Hoc On Demand Distance Vector Routing (AODV). Because both AODV and DSR do not use location information in their route discovery, they are likely to suffer from added overheads in form of latency. This is so because during route discovery, AODV and DSR initiate the process for the first time without prior route information. Oppositely, proactive routing protocols have less latency because of constant updates of routing information.

Another key common feature between AODV and DSR is the flooding of data packets during route discovery process. Proposition of LAR is that utilization of location information improves the performance of such reactive protocols. Although flooding as design feature is maintained in LAR, introduction of location information as an algorithmic feature helps in the reduction of latency and the overall packets throughput. For example, figure 4 represents the concept of flooding in either AODV or DSR. Assuming that node S is the sender, it broadcasts a route discovery request to all of its neighbors [23].

If a neighboring node does not recognize the routing information in the route request packet, it discards the it. Otherwise, it forwards to the next neighbor. Every node makes the decision either to forward or drop a route discovery packet by comparing the intended destination information with their identifiers. However, sequence numbers are used to avoid redundancy. Therefore, a node is expected to conduct route request forwarding once.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

In figure 4, sender node S broadcasts a route request message to nodes A, B, and C. But since node A has no neighbors apart from S, it discards the message. Nodes B and C forward the request message to E and F respectively. Similarly, node E discards the message because it is the dead end. However, nodes B and C forwards the message to node F. However, node F retains the message that is received first while discarding the other message that comes second. Eventually, node F forwards the rout request to destination node D.

Node D then sends a reply to source node S based on the route through which it received the request message—request message retains the nodal routing information as moves from the source/sender node S to destination node D.

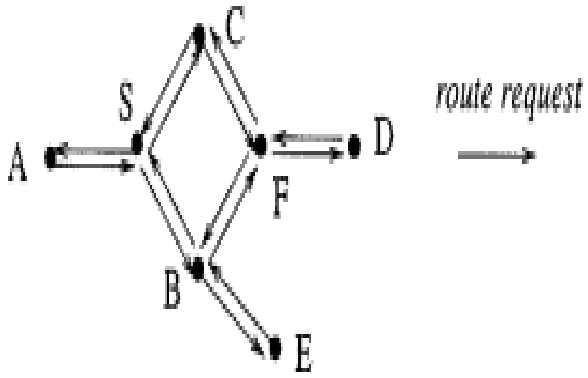


Figure 4: Example of Flooding in either AODV or DSR

If the destination node fails to receive the request message either due to link error, a timer is used to determine when a resend is needed. The main purpose of LAR is to diminish the number of nodes that receive the routing packet. By doing so, it creates a partial flooding. Utilization of location information by the LAR in route discovery is intended to introduce elimination of certain nodes that are not within either the expected or request zones—refer to figures 5 and 6.

Location information can be obtained through Global Positioning System (GPS). The key assumption of LAR is that the source node knows the maximum mobility speed of the destination node. In figure 5, the expected zone is the area determined by multiplying the speed of the destination node and the time between the request messages. The results from the speed and the time difference offers the radius of the expected area.

Assuming that the destination node S has a maximum speed V, at time t_0 (previous time), then at time t_1, the expected zone is the area determined by radius $v(t_1-t_0)$. Although comparative simulation studies suggest that LAR outperforms normal flooding algorithms such as AODV and DSR, it still experiences latency due to partial flooding. In other words, LAR uses minimized flooding in its routing despite introduction of location information. The presence of partial flooding retains some level of latency.

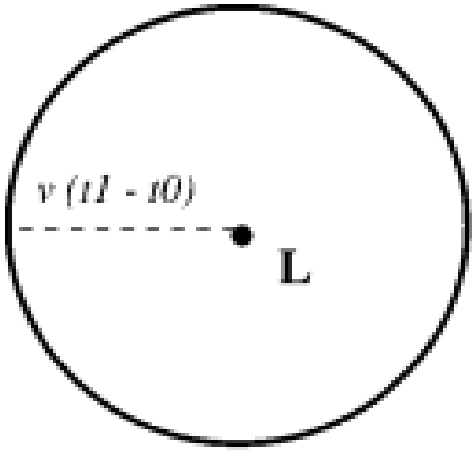


Figure 5: Expected zone at t_1

With inclusion of more routing information of the destination node, the expected zone reduces in size, hence reducing the amount of route request time. For example, if the source node S knows that direction of movement of destination node S is north, the expected zone in figure 5 can be reduced into half. Figure 6 represents reduction of the expected zone by half. More detailed workings of LAR can be found in paper [23]. However, LAR relies on two principles in its routing. That is, the expected zone and the request zone. The purpose of this paper is not to explain LAR in details but rather to offer an alternative to partial flooding.

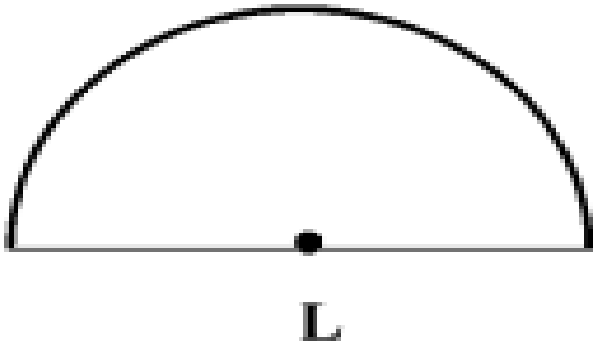


Figure 6: Expected zone reduced by half

5. Proactive Location-Aided (P-LAR)

Conventional LAR uses partial flooding of the expected and request zones based on location information of the destination node. Methods used in determining both the expected and request zones has been described in the previous sections of this article—refer to figures 5 and 6.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

Figure 7 (a) demonstrates partial flooding on the expected and request zones in LAR Scheme 1. In figure 7 (a), source node S is outside the request zone.

The expected zone is defined by (Xd, Yd) coordinates and radius R. The Request zone is the rectangular area defined by A, B, C, S. Because node J—defined by (Xj,Yj) coordinates, is outside the request zone, it does not receive request messages broadcasted by source node S. In figure 7 (b), the source node S—defined by (Xs,Ys) coordinates, is within the request zone. As noted earlier, radius R is determined through multiplication of estimated nodal speed v with the difference between prior time t_0 and the current time t_1; (t_1-t_0).

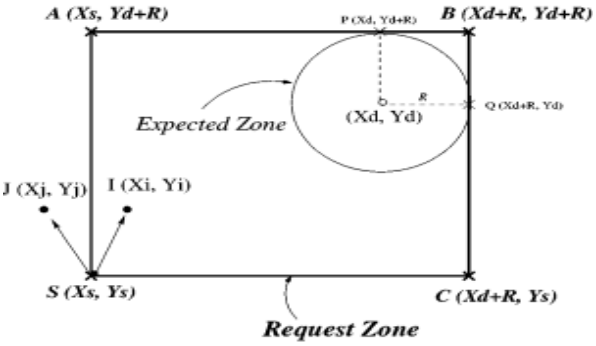


Figure 7: Expected Zone with and without the Source Node S

In this study, we propose introduction of proactivity in LAR route discovery process. The argument is that constant update of location information among network nodes are will eliminate the need for partial flooding. Topological routing tables are used to store three key location information of the nodes within the transmission range.

The two-location information are the Angle of Arrival (AoA), the distance from the source node and the expected direction of movement. In the new proactive LAR (P-LAR), the source node receives constant updates of location information of all nodes within a request zone. Proactivity within the source nodes speeds up route discovery process because of the updated location information in the network topological routing tables.

Position = initial position+ initial velocity * time + 1/2 * acceleration * (time)^2.

x = position

x0 = preliminary position

v0 = preliminary velocity

t = time

a = acceleration

Although proactivity may increase other routing overhead such as power, latency is greatly reduced. Minimization of latency results into better general data packet throughput. Augmenting general data packets throughput increases the performance of a MANET network. Better performance improves the general performance of MANETs. Because MANETs are used in areas whose infrastructure is damaged or missing altogether, reliable performance is vital. In our proposal, the location information of the destination node is composed of the Angle of Arrival (AoA) and the distance from the source node. These two parameters calculable based on the coordinates of the nodes.

Assume, for example, that in figure 7, the coordinates subset of destination node D is (x_0, y_0) at time t_0. Due to its movement at time t_1, the coordinates subset for the destination node D is (x_1, y_1). Therefore, the distance covered due to the movement of the node is;

$$d = \sqrt{((x_1-x_0)\ \ ^2+(y_1-y_0)\ ^2)} \qquad (1)$$

Else, if the movement speed of destination node D is a constant v, then the distance d is determinable based on the following formulae;

$$d = v\ (t_1-t_0) \qquad (2)$$

$$d = x_0 + v_0t + \ a*t2/2 \qquad (3)$$

Similarly, the Angle of Arrival (AoA) is determined based on the following formula;

$$\theta = \tan\ ^(-1)\ \ [\!\![((y1 - y0)\ \ / ((x1 - x0)\ )]\!\!] \qquad (4)$$

The above location information is included in the algorithm below and run on C++ based simulator. It is critical to note, however, that the code snippet included below is only a representation of a single modular function to enable proactivity in LAR. The code below was only tested in OMNET++ simulator. For simplicity purposes, integer is used for both the distance and the angle of arrival (AoA).

Due to proactivity, the following nodes formula— (5), must be applicable to maintain continuousness of the location updates.

V_n is the nth node during the proactivity process.

N is the number of nodes in the network.

Therefore, to assure constant location information update;

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

$$\sum n = 1NVn = 0 \qquad (5)$$

Rendering of the code on OmNET++;

```
network P-LAR
{
   parameters:
      int numHosts;
     int nodeDist;
    int  AoA;
      @display (nodeDist, AoA);
      @display("bgb=650,450");
      @\displaystyle\sum\limits_{n=1}^N V_n = 0;
   submodules:
      visualizer:  <default("IntegratedCanvasVisualizer")>
like IIntegratedVisualizer if hasVisualizer() {
         parameters:
            @display("p=100,300;is=s");
      }
      configurator: Ipv4NetworkConfigurator {
         parameters:
            @display("p=100,100;is=s");
      }
      radioMedium: UnitDiskRadioMedium {
         parameters:
            @display("p=100,200;is=s");
      }
      host[numHosts]: AdhocHost {
         parameters:
            @display("r=,,#707070;p=300,200");
      }
}
```

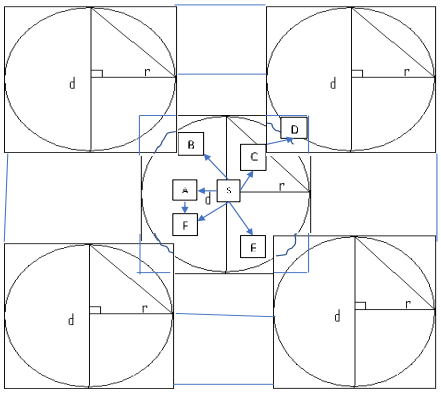


Figure 8: A theorical representation of initial routing instance in the P-LAR

Reference Source Node s Actions:

-Physical location information is included in the topological routing table of node s

-Physical information includes the Angle of Arrival (AoA) and the relative distance of the destination node s and the

-Relative distance and the AoA is used to determine the presence of the destination node in the request zone as defined by radius r in figure 8.

-If the request zone does not cover the topological distance and AoA, the destination node is outside the request zone, hence removal of the information from the network topological routing table.

- Route dismissal from the network topological routing table is initiated through equation 5

- Else, a network topological route discovery message is relayed to the source node

-Otherwise, the data packet with the physical location information of node as identified in network topological routing table, is successfully delivered to the appropriate destination

6. Evaluation of Performance of P-LAR

Because this study was designed to gauge the effects of proactivity inclusion in LAR, OMNET++ was used to simulate the LAR Scheme 1 and the proposed new Proactive LAR Scheme 1—P-LAR. Results on basic scalar parameters—network throughout, packets drop rate, latency,

and the data packets successfully received, were analyzed on RapidMiner to produce graphical illustrations. It is important to emphasize that the graphs presented herein are summaries of three iterations of simulations.

As noted above, location information included in the performance evaluation was limited to the distance of the source node, Angle of Arrival (AoA), and the general estimation of direction of the destination node. Simulation model required 10,100 and 1000 nodes for test runs. However, the graphs included in this study are summaries of average results of these three test runs. Table 1 describes the summary of parameters, variables and their associated metric indicators. Simulation environment is largely derived from []. All packets transmission is assumed to occur at Transmission Control Protocol/Internet Protocol (TCP/IP) layer of network.

Table 1: Parameters, Variables and Indicators

Parameters	Parameter Value	Simulation Variables	Variable Indicators/Values
Packet size	1,500 bytes	Network Throughput	Quantity of data packets
Nodes	10,100,1000	Drop Rate	Amount of Data Packets dropped/lost in an iteration
Simulation Time	100 seconds	Network Latency	Transmission time in seconds
Number of iterations	3	Data Received	Amount of data packets received by destination node

Each test run duration was 100 seconds. The area of movement was confined to 500m X 500 m. Uniform distribution of the coordinates was used to determine variable x and y. Initial assumption is that nodes do not move while subsequent assumption is that all nodes move at a constant average speed of *v*. General direction of the nodes movement is determinable through Global Positioning System (GPS). Another simulation assumption is that the location error is negligible. Since all nodes are assumed stationary during the simulation event:

time $t_0=t_1$; therefore, time difference = 0;

also; velocity $v_0=v_1$, therefore, distance d = 0;

Network throughput (PT) is key this study because it represents the overall performance of the proposed P-LAR compared to LAR Scheme 1. This study measures network throughput as the key basic scalar parameter in the simulation because it the only performance metric that can be influenced by variations in latency, packet loss, network congestion, and jitter. However, this study derives simulation parameters from the latter four metrics to simulate latency, packets delivery ratio and the number of packets dropped.

Figure 9 represents the simulation results for data packets throughput (PT) or network throughput (NT) simulation results between the conventional LAR Scheme 1 and the proposed Proactive LAR Scheme 1 (P-LAR). Data packets throughput in this simulation refers to the quantity of data successfully sent and successfully received within a specified period of time.

It can also refer at the rate at which packets are successfully received at the destination node. That is, the result of dividing the number of the packets received (PR) at the destination node with the number of packets successfully transmitted (PT) at the source node. In this simulation, the units used in measuring this quantity are expressed in Megabits per second (Mbps). The following formula was used to calculate network throughput:

$$NT = \frac{PR}{PT} \qquad (6)$$

Figure 9 represents the general average trend of data packet throughput of three simulations runs—10,100 and 1000 nodes. Clearly, there is a general variation in the network throughput between the Proactive LAR (P-LAR) and the traditional LAR Scheme 1. However, at around 90th second, P-LAR drops sharply relative to LAR Scheme 1. At around 95th second, P-LAR rises to supersede Network throughput of LAR Scheme 1.

Although the performance of P-LAR looks steadily better than conventional LAR Scheme 1, the variation in ultimate data packets throughput can be phenomenal after long period of time. Because this study is only concerned demonstrating fundamental effects of inclusion of proactive algorithmic elements in LAR Scheme 1, further studies can be conducted to derive the appropriate extrapolation formula that can be used in determining the variations in network throughput at $t_0 \ldots t_n$. The general overall of p-LAR is represented by the dotted line.

On average, P-LAR successfully transmits some 1665 data packets per second (1665.17/sec) while traditional LAR successfully transmits some 1643 data packets per second (1643.28/sec.). The variance in transmission success (DT) is (1665.179-1643.28)/sec.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

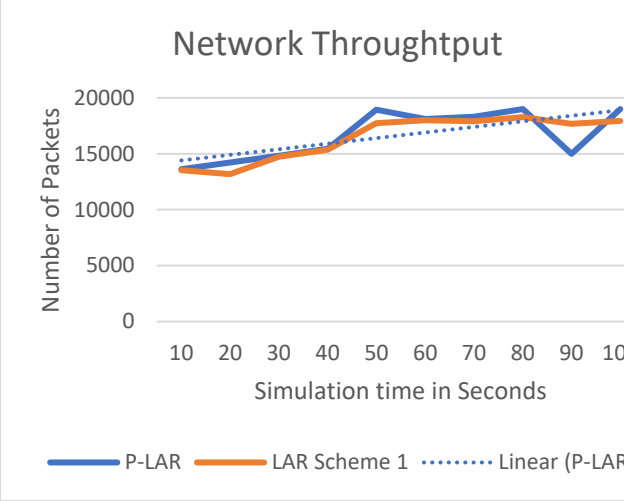


Figure 9: Network Throughput of LAR Scheme 1 and P-LAR

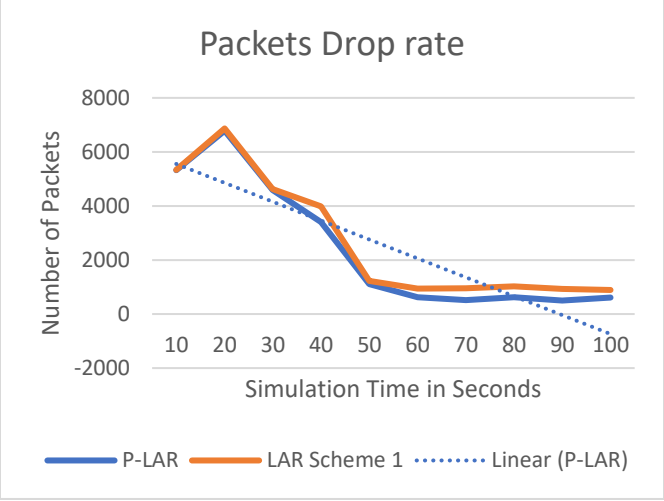


Figure 10: Data Packets Drop rate

After simulating the network throughput, this study examined variations in the number of data packets lost/dropped. It should be noted that the rate of data loss affects the general final network throughput. Therefore, it is only reasonable that P-LAR outperform the traditional LAR Scheme 1. The rate of data packets (PL) droppage or loss is calculated by dividing the number of data packets successfully transmitted (PT) at source node with the data packets successfully received (PR) at the destination. The following formula was used in determining the rate of data droppage during the simulation period:

$$\text{PL} = \frac{PT}{PR} \qquad (7)$$

After simulation run, the graphical representation shows that, as expected, inclusion of proactive algorithmic elements in LAR Scheme 1, lowers the number of packets lost during route discovery and data packets transmission. In figure 10, the proactive LAR steadily and consistently experienced lower data packets drop rates. As the P-LAR trend line indicates, the proactive algorithmic routing element lowers the packets drop rates.

However, up to around 30th second, P-LAR and the conventional LAR Scheme 1 experience virtually similar levels of data packets drop rates. After the 30th second point the proactive enhanced consistently outperforms the traditional LAR Scheme 1. On average, P-LAR drops some 240 data packets per second (240.81/sec.) while traditional LAR Scheme 1 drops some 268 per second (268/sec.). The differential drop rate (DD) per second is approximately (240-268)/sec.

Another performance parameter evaluated in this study is latency. Network latency (NL) is the measure of the amount of time a data packets takes to successfully move from the source node to the destination node. Latency in this study is the measure of the round trip taken by a data packet from the source node to destination node and back. Usually, maximum latency is determined by dividing the desired size of the packet size (DP) with the with maximum network throughput (NT). The following formula is used to determine latency:

$$NL = \frac{DP}{NT} \qquad (8)$$

Figure 11 represents latency results of both the P-LAR and the traditional LAR Scheme 1. The graph indicates that, as expected, the overall latency of P-LAR is relatively lower throughout the simulation instances. The results are consistent with the rest of measured basic scalar parameters. Lower latency in P-LAR means that a data packet takes less time for a roundtrip during route discovery and packet transmission.

P-LAR experiences average latency of 5.64E-03 seconds, while traditional LAR Scheme 1 experiences 6.36E-03 seconds of latency. The latency variation is (5.64E-03-6.36E-03) seconds. Although the -7.20E-04 seconds in latency variation may not seem as a big difference, it can greatly influence the overall performance of the network over a longer period of time. Moreover, such small difference in latency can cause a huge difference when transmitting large sizes of data packets.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 193-204, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1004

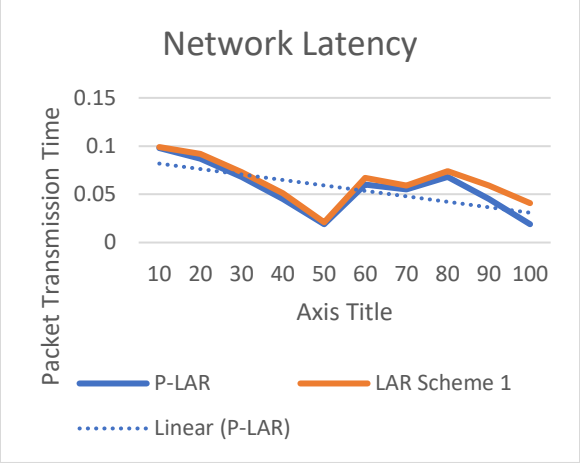


Figure 11: Network Latency

Average data packets received successfully (APR) measures the quantity of data packets effectively transmitted from source node to the destination node in a given period of time. APR is calculated by dividing the number of data packets successfully transmitted with time (T) taken to transmit the packets. The following formula is used in determining the average data packets transmitted:

$$APR = \frac{PR}{T} \qquad (9)$$

Similar to other three basic scalar performance parameters, the quantity of average data packets received is relatively better in P-LAR than the conventional LAR Scheme 1. This result is in line with network throughput, packets drop rate and latency.

Because the network throughput in P-LAR is relatively higher than conventional LAR Scheme 1, it means that it experiences less packet loss, better latency and higher average data packets transmission. Figure 12 represents the graphical representation of the performance of P-LAR and the conventional LAR. The average data packets successfully received variant is (1167 - 1101.11) packets. This translates to transmission of more 66 data packets by the P-LAR compared to traditional LAR Scheme 1. However, the difference in average data packets received seem to congregate after the 100th second.

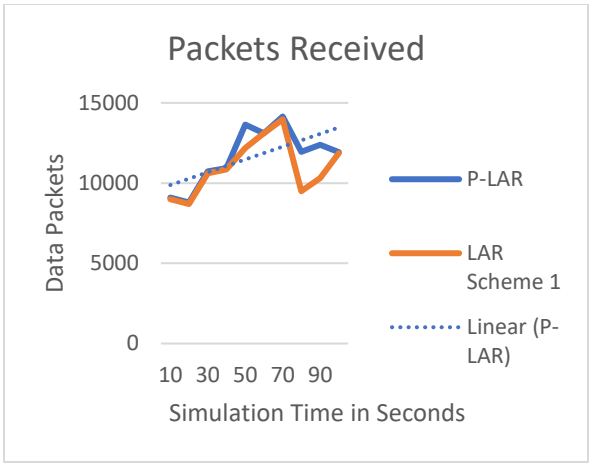


Figure 12: Data Packets Received

7. Conclusion

This study examines routing performance effects of proactive algorithmic feature in conventional LAR Scheme 1. The proposed version of LAR Scheme 1 is called Proactive LAR Scheme 1 (P-LAR). The simulation was carried out on OMNET++ while data analysis was done on RapidMiner. Simulation results of the new algorithm (P-LAR) suggest that inclusion of the proactive algorithmic element augments the performance on network throughput, while minimizing packet loss and latency. Besides the proactive algorithmic element enhances the amount of average data packets reception. This performance variation is attributable to the ability of proactivity routing algorithm to reduce the size of the request zone, hence minimized flooding during route discovery process.

Because this study focused on elementary scalar parameters, we recommend that further studies be carried to gauge the routing performance effects of proactive algorithm on jitter and power consumption. Moreover, it is important to extend this research to include the LAR Scheme 2. We also recommend examination of the effects of proactive algorithm on routing performance other location-aided protocols such as the Greedy Perimeter Stateless Routing (GPSR).

References

[1] D. Ismail and M. Ja'afar, "Mobile ad hoc network overview", *2007 Asia-Pacific Conference on Applied Electromagnetics*, 2007. Available: 10.1109/apace.2007.4603864 [Accessed 9 September 2020].

[2] K. Weniger and M. Zitterbart, "Address autoconfiguration in mobile ad hoc.netvvorks: current approaches and future directions", *IEEE Network*, vol. 18, no. 4, pp. 6-11, 2004. Available: 10.1109/mnet.2004.1316754 [Accessed 9 September 2020].

[3] S. Jadhav, A. Kulkarni and R. Menon, "Mobile Ad-Hoc Network (MANET) for disaster management", *2014 Eleventh International Conference on Wireless and Optical Communications Networks (WOCN)*, 2014.Available: 10.1109/wocn.2014.6923074 [Accessed 9 September 2020].

[4] R. Thiagarajan and M. Moorthi, "Efficient routing protocols for mobile ad hoc network", *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2017. Available: 10.1109/aeeicb.2017.7972346 [Accessed 9 September 2020].

[5] W. Liu and H. Song, "Research and implementation of mobile ad hoc network emulation system", *Proceedings 22nd International Conference on Distributed Computing Systems Workshops*. Available: 10.1109/icdcsw.2002.1030858 [Accessed 9 September 2020].

[6] S. Meshram and P. Dorge, "Design and performance analysis of mobile Ad hoc network with reactive routing protocols", *2017 International Conference on Communication and Signal Processing (ICCSP)*, 2017. Available: 10.1109/iccsp.2017.8286396 [Accessed 9 September 2020].

[7] N. Milanovic, M. Malek, A. Davidson and V. Milutinovic, "Routing and security in mobile ad hoc networks", *Computer*, vol. 37, no. 2, pp. 61-65, 2004. Available: 10.1109/mc.2004.1266297 [Accessed 9 September 2020].

[8] H. Moudni, M. Er-rouidi, H. Mouncif and B. El Hadadi, "Secure routing protocols for mobile ad hoc networks", *2016 International Conference on Information Technology for Organizations Development (IT4OD)*, 2016. Available: 10.1109/it4od.2016.7479295 [Accessed 9 September 2020].

[9] P. Lalwani, S. Silakari and P. Shukla, "Optimized and Executive Survey on Mobile Ad-hoc Network", *2012 International Symposium on Cloud and Services Computing*, 2012. Available: 10.1109/iscos.2012.37 [Accessed 10 September 2020].

[10] M. Athulya and V. Sheeba, "Security in Mobile Ad-Hoc Networks", *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*, 2012. Available: 10.1109/icccnt.2012.6396047 [Accessed 10 September 2020].

[11] L. Zhitang and S. Shudong, "A Secure Routing Protocol for Mobile Ad hoc Networks", *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*, 2007. Available: 10.1109/icis.2007.43 [Accessed 10 September 2020].

[12] A. Junnarkar and A. Bagwan, "Efficient algorithm and study of QoS-aware mobile Ad hoc network methods", *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017. Available: 10.1109/icoei.2017.8300842 [Accessed 10 September 2020].

[13] S. Sharmila and T. Shanthi, "A survey on wireless ad hoc network: Issues and implementation", *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, 2016. Available: 10.1109/icetets.2016.7603071 [Accessed 10 September 2020].

[14] N. Enneya, K. Oudidi and M. Elkoutbi, "Network Mobility in Ad hoc Networks", *2008 International Conference on Computer and Communication Engineering*, 2008. Available: 10.1109/iccce.2008.4580751 [Accessed 10 September 2020].

[15] D. Niu, Y. Zhang, Y. Zhao and M. Yang, "Research on Routing Protocols in Ad Hoc Networks", *2009 International Conference on Wireless Networks and Information Systems*, 2009. Available: 10.1109/wnis.2009.36 [Accessed 10 September 2020].

[16] R. Gupta, N. Krishnamurthi, U. Wang, T. Tamminedi and M. Gerla, "Routing in Mobile Ad-Hoc Networks Using Social Tie Strengths and Mobility Plans", *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017. Available: 10.1109/wcnc.2017.7925620 [Accessed 10 September 2020].

[17] K. Rinku, M. Manish and B. Megha, "Energy efficient routing in mobile Ad-hoc network", *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 2017. Available: 10.1109/ssps.2017.8071563 [Accessed 10 September 2020].

[18] K. Wu and J. Harms, "Load-sensitive routing for mobile ad hoc networks", *Proceedings Tenth International Conference on Computer Communications and Networks (Cat. No.01EX495)*. Available: 10.1109/icccn.2001.956319 [Accessed 10 September 2020].

[19] V. Hnatyshin, M. Ahmed, R. Cocco and D. Urbano, "A comparative study of location aided routing protocols for MANET", *2011 IFIP Wireless Days (WD)*, 2011. Available: 10.1109/wd.2011.6098169 [Accessed 10 September 2020].

[20] M. Nabil, A. Hajami and A. Haqiq, "Improvement of location aided routing protocol in Vehicular Ad Hoc Networks on highway", *2015 5th World Congress on Information and Communication Technologies (WICT)*, 2015. Available: 10.1109/wict.2015.7489644 [Accessed 10 September 2020].

[21] N. Wang and S. Wang, "An Efficient Location-Aided Routing Protocol for Mobile Ad Hoc Networks", *11th International Conference on Parallel and Distributed Systems (ICPADS'05)*, 2019. Available: 10.1109/icpads.2005.82 [Accessed 10 September 2020].

[22] J. Meng, H. Wu, H. Tang and X. Qian, "An Adaptive Strategy for Location-Aided Routing Protocol in Vehicular Ad Hoc Networks", *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 2013. Available: 10.1109/imis.2013.75 [Accessed 10 September 2020].

[23] Y. Ko and N. Vaidya, *Wireless Networks*, vol. 6, no. 4, pp. 307-321, 2000. Available: 10.1023/a:1019106118419 [Accessed 10 September 2020].

Authors

Mutuma Ichaba

I have over three years of working with e-learning systems. Currently, I am pursuing a Ph.D in Information Systems, specializing in Mobile Ad Hoc Networks (MANETs) at KCA University in Kenya. I am particularly interested in the use of Mobile Ad Hoc Networks (MANETs) in the creation of virtual classrooms for the digitally disadvantaged groups such as slum dwellers and pastoralist communities in the developing countries.

Prof. Felix Musau

He an expert in Computer Science and Information Technology where he worked over years in both Technical and Academic fields. Professor Felix Musau graduated with a Doctor of Philosophy degree and Master of Science in Computer Science and Technology, from the School of computing and Information Engineering, Central South University, P.R CHINA in 2012.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 205-212, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1005

Machine Learning Prediction Models for Postpartum Depression, a Review of Literature

George Kimwomi
Institute of Computing and Informatics
Technical university of Mombasa
Mombasa, Kenya

Mvurya Mgala
Institute of Computing and Informatics
Technical university of Mombasa
Mombasa, Kenya

Fullgence Mwakondo
Institute of Computing and Informatics
Technical university of Mombasa
Mombasa, Kenya

Pamela Kimeto
School of Medicine and Health Sciences
Kabarak University
Nakuru, Kenya

Abstract: Postpartum depression is a medical condition which continue to affect many mothers after delivery even though the disease can be prevented. It consequently exposes mothers and family members to illness and even death. Families, governments and other stakeholders incur heavy expenditure in the management of the disease. Research studies have been done to develop machine learning models for prediction of mothers at risk of postpartum depression during pregnancy for preventive measures. This paper presents a literature review of the machine learning prediction models which have been developed for the condition with specific focus on feature selection methods, algorithms used and the resulting performance. Literature review was done with google scholar integrated to an online institutional account for e-resources from e-databases accessed by subscription or free access. Inclusion involved all articles with the key words "machine learning, prediction model, postpartum depression" in the articles dated from 2018 to 2022 and sorted by relevance. A total of 3430 articles were listed while only 17 which were accessible with full text were eligible and therefore selected for the study. Analyzes were done using Microsoft Excel and descriptive analysis. Findings and conclusions will inform scientists on the status of research in the area to guide new studies, and inform the market on the potential benefits of integrating machine learning models in their systems.

Keywords*:* Machine learning, prediction, model, postpartum depression

1. INTRODUCTION

1.1 Machine Learning

Advancement in computing technology has given rise to alternative methods of operation across various industries leading to improvement on service delivery. A notable example is machine learning (ML) technology which creates artificial intelligence in computer systems to aid them in solving new problems. ML is a technology which makes computers to study and simulate human activities so as to acquire artificial intelligence that enables them to learn from experience using historical data and apply the knowledge acquired to solve similar problems without explicit reprogramming[1]. Information is extracted from complex datasets which enable computers to make intelligent decisions which improves their performance. It is an emerging technology triggered by improved methods of capture and storage of data following the advancement of data management techniques. ML technique is derived from the methods traditionally used to analyze data inputs and extract information which include mathematics, statistics, data mining, optimization and artificial intelligence[2].

According to Lai *et al*, three different methods of ML can be used which includes supervised learning which uses labelled datasets to develop a model, the unsupervised learning which can discover data patterns automatically from unlabeled dataset based on a given criteria, and reinforcement learning which also uses unlabeled dataset whereby learning is achieved through experience from interaction with the environment. A range of different algorithms can be used to train ML models and select the best performing model[3][4]. The ML process entails collection and preparation of training and test datasets for model development. By the use of specialized ML tools, training is done on a training dataset to create a preliminary model where the pattern between the input and output data is established. The resulting model is tested with unlabeled dataset purposely set aside for testing the model to confirm its accuracy in predicting outcome from new data input.

ML models can be used in various domains to predict events or conditions for awareness so as to prevent or plan on how to counter them. It can specifically be used to support medical personnel in the prediction postpartum depression for pregnant mothers using antenatal data to improve management of the condition. This study could help identify gaps in ML prediction models for postpartum depression and serve as background for development of new prediction models.

1.2 Feature selection and ML algorithms

The choice of the feature selection method and machine learning algorithm used in model development is an important stage in machine learning which will determine the reliability of the developed model. Features in ML are the inputs to a model while the output is described as the response or independent variable for a model [5]. A research problem could have a wide range of input characteristics while only a given fraction of the variables is significant in predicting the target variable. Model development should thus be an elaborate process involving careful feature selection which can accurately predict the correct outcome. This is achieved through different feature selection procedures which are compared during model training to select a method which can produce the most optimal features. The feature selection methods which have been used in healthcare prediction

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

models include random forest[6], sequential feature selection[3], optimizer[7] and SelectKBest[8]. Expert judgement can also be used by experienced personnel to select features but the automated methods have proved to give better performance[4]. different machine learning algorithms are also suitable for specific kinds of problems and should therefore be carefully selected when developing models. Algorithms such as support vector machines, decision tree, regression and Naïve Bayes are suited for supervised classification models while K-means is suited to unsupervised classification[9]. Whereas a range of algorithms could be suited for a certain kind of problem, model development should involve trial of several selected algorithms for comparison in order to select the most the best performing choice. Feature selection methods should also be tried with different ML algorithms in different environments during training to identify the combination those circumstances.

1.3 Postpartum depression

Good health and well-being is one of the 17 Sustainable Development Goals (SDGs) of the United Nations(UN) which the body seeks to achieve by the year 2030[10]. Among these health concerns is postpartum depression (PPD) which affects about 10% to 20% mothers after delivery, and could by extension have serious effects on the new born baby and other family members[4]. The prevalence rates are not collectively accepted as a reflection of the actual rate since different studies have shown varying rates while there is belief that many cases are not reported. PPD is a serious mental health which can affect mothers for up to one year after delivery[3]. Victims of the condition exhibit associated signals such as sleep disorder, irritability, anxiety, and stress. Worse cases could include intents to murder or commit suicide which can be actualized if proper intervention is not given to the victims in good time. Its effect on children can continue beyond childhood with problems such as weight loss, mental retardation, poor physical growth and other vulnerabilities[11]. Medical personnel use self-reporting questionnaire tools and hospitals personnel expertise to predict the risk of PPD as there are no laboratory methods for the prediction[4][12]. W. Zhang *et al* identified a range of antenatal features like demographics, psychology, diagnoses and client environment as essential characteristics for the prediction of PPD during pregnancy. ML can be integrated to data management systems and use such features to develop systems which can reliably predict mothers at risk of PPD during pregnancy for better management of the condition. The aim of this study was to analyze the feature selection methods and ML algorithms used in the development of ML prediction models for PPD and the resulting performance.

2. RELATED STUDIES

A survey of past review studies on the machine learning models was necessary to reveal the trend in feature selection methods, algorithms used and the resulting performance as summarized in table 1. [13] Carried a scoping review using Arksey and O'Malley frameworks from health and information technology databases covering a period of 12 years. Supervised machine learning technique was used by the entire publications covered while the different algorithms used produced varying performing outcomes with the Area Under the receiver operating characteristic Curve (AUC) ranging from 0.78 to 0.93. The studies did not report on the feature selection procedures used but revealed the potential of using ML technique in the prediction of PPD. Repeated modelling of the different algorithms as concluded by the author could help identify the most suitable combination of feature selection procedures and ML algorithms alongside other parameters to create better performing models. Another literature review by[14] using the PubMed and Embase databases found support vector machine to be the most popular algorithm while all the studies achieved a AUC of over 0.7 which was considered as an acceptable performance in the prediction of PPD. Feature selection methods were not reported but further studies were recommended to advise how such models could be applied in actual practice to support healthcare predictions. Another review by [15] found Bayes Net classifier to be the best performing model (AUC=0.93) compared to support vector machine (SVM), decision trees and neural networks, among others. The feature selection procedure used was not reported. The author concluded that the findings of the literature review was an opener to further research which is a recommendation for further research.

These studies revealed that PPD prediction models developed achieved over 70% accuracy which was considered satisfactory for implementation of the technology in health management. Support vector machine was the most frequently used algorithm while Bayes Net classifier and logistic regression produced the most accurate performance (AUC of 0.93). The other algorithms which achieved a 70% accuracy or higher model performance are Random forest, XGBoost and logistic regression which revealed the potential of ML in the prediction of PPD. A limited number of review studies were found which applied machine leaning to predict the condition. The few review studies also missed to report on feature selection procedures used for the past prediction models. These gaps justified the need for another review to provide missing information and report on progress achieved from recent research studies.

Author / year	**Feature selection Methods**	**Most used Algorithms**	**Performance**
[15], 2020	Not reported	Bayes Net classifier	AUC=0.93
[13], 2021	Not reported	logistic regression	AUC= 0.93
[14], 2022	Not reported	SVM	AUC >0.70

Table 1: Summary of feature selection methods and ML algorithms from related studies

3. METHODOLOGY

3.1 Search criteria

A web-based search with google scholar integrated to an institutional online e-resources account was used to retrieve primary research publications accessible through subscription or open access. The key search words used (model AND prediction AND depression postpartum OR postnatal "machine learning") were formulated from the research objectives and Boolean operators to retrieve required articles. The search which was done in the month

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

of April to early May 2022 also applied filters to select primary articles and limit target period to the years 2018 to 2022 which were sorted by relevance as shown in figure 1.

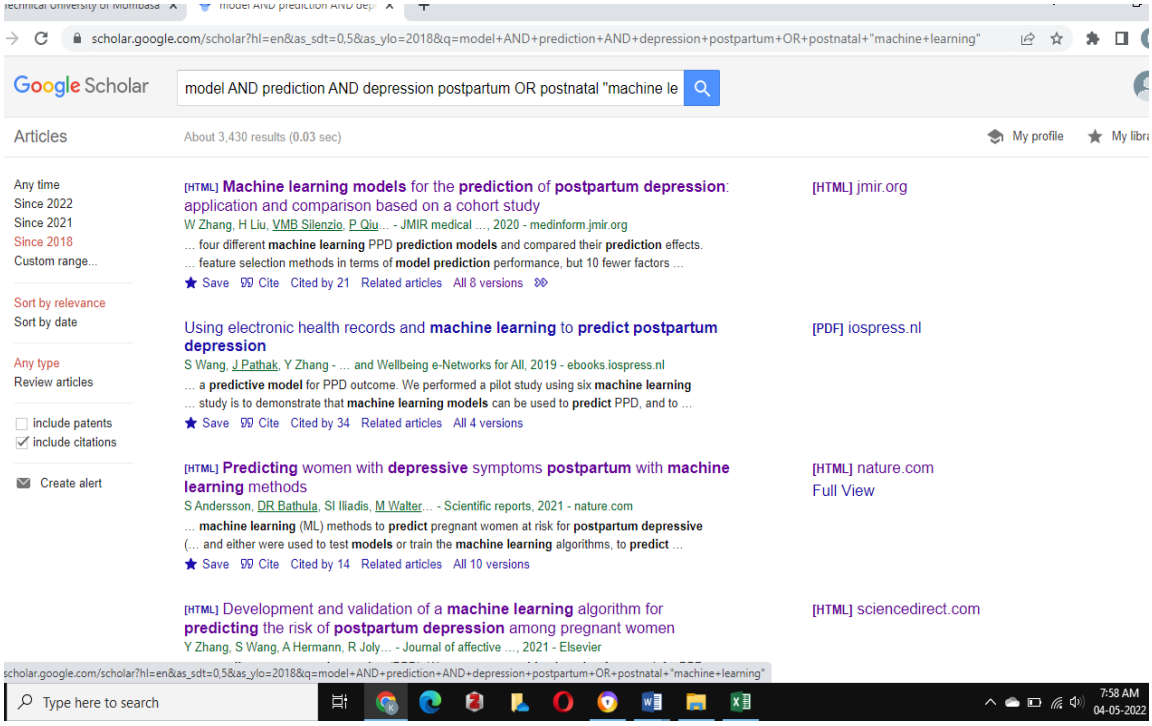


Figure 1: A cross-section of the web search result for eligible publications

3.2 Inclusion and exclusion

Articles for inclusion were primary publications that used antenatal and postnatal data in modelling and specifically for prediction of postpartum depression. The models should have been developed using machine learning technique and the publication period dated from the years 2018 and 2022. The studies for inclusion must have used medical records that originated from hospital sources for the analysis. Articles using secondary publications or those which their full text was not accessible were excluded. Articles on detection of depression which did not focus on future prediction of postpartum depression were also excluded. Retrieval and analysis of the articles was done from Aril to early May 2022. The articles retrieved in the initial search were screened for eligibility by two researchers who evaluated the titles and abstracts. Duplicate and irrelevant articles, and studies which did not use data from hospital medical records sources or which were not specifically focused on future prediction of PPD were excluded.

3.3 Data collection and analysis

Feature selection procedures, ML algorithms and performance of the developed models were extracted by reading the abstracts, methodology, results and conclusion sections of the eligible articles which were captured in table 2. A narrative synthesis of the data was done for the articles in regards to the research criteria. The analysis revealed the findings which supported the conclusions made which could help scientists know the status of research in the area and inform the market about the potential benefits of integrating machine learning models in their systems. This was also important for future studies in the development of new ML models for prediction of PPD to fill gaps identified.

4. RESULTS

4.1 Search process results

A total of 3,430 articles were identified from the initial search process which were subjected to screening out of which 3173 consisting of duplicates, non-articles and irrelevant articles were excluded. The remaining 255 articles were evaluated for eligibility out of which 239 articles were excluded due to missing of the full text and failing to use data from hospital record sources. A total of 16 articles which were found to be eligible were included for the study as illustrated in figure 2.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

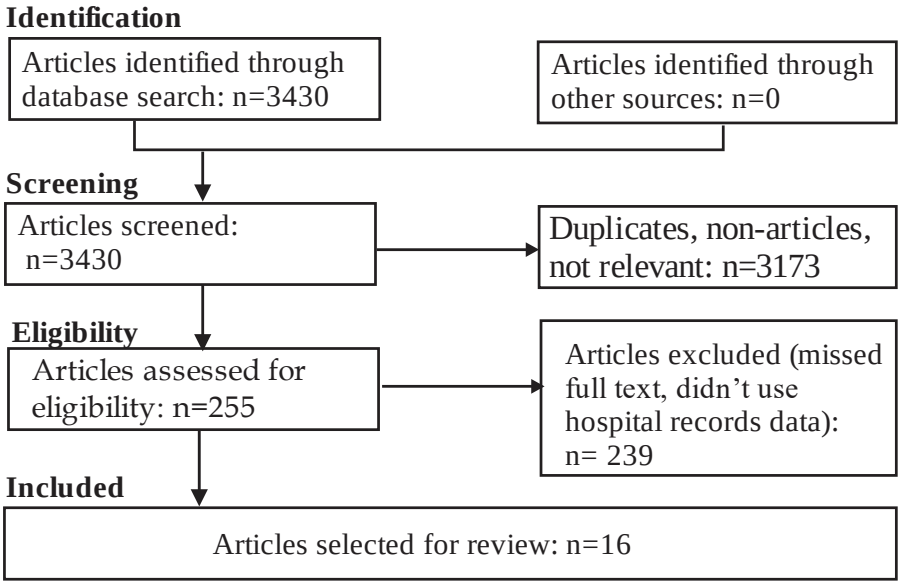


Figure 2: Search process for eligible publications

Table 2: Summary of result for selected studies, feature selection methods and ML algorithms used

SN	Author / Year	Title of publication	Feature selection method	ML Algorithm used	AUC
1	[3], 2021	Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women	SFS	Logistic regression with L2 regularization	0.937
2	[16], 2021	Recommender System for Postpartum Depression Monitoring based on Sentiment Analysis	NLP	Text mining	0.88
3	[17], 2021	Predicting women with depressive symptoms postpartum with machine learning methods	Gini Importance or MDI	Extremely randomized trees	0.73
4	[8], 2021	An in-depth analysis of machine learning approaches to predict depression	SelectKBest	AdaBoost	0.96
5	[6], 2021	Predicting Individuals Mental Health Status in Kenya using Machine Learning Methods	RF Classifier	Voting-Ensemble	0.85
6	[17], 2021	A Community Based Study for Early Detection of Postpartum Depression using Improved Data Mining Techniques	J48 algorithm	Adaptive Boosting Collaboration	0.94
7	[18], 2021	Estimation of postpartum depression risk from electronic health records using machine learning	SHAP	Gradient tree boosting algorithm	0.844
8	[19], 2021	Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study	gradient-boosted decision tree	XGBoost	0.712
9	[20], 2021	Machine Learning Models for the Prediction of Postpartum Depression: Application and Comparison Based on a Cohort Study	FFS-RF	SVM	0.78
10	[21], 2020	Depression Detection using Machine Learning	Vectorization	Naïve Bayes	0.936
11	[22], 2020	Data-Driven Insights towards Risk Assessment of Postpartum Depression.	ReliefF expRank	RF	0.75
12	[23], 2020	Machine learning-based predictive modeling of postpartum depression	Relief algorithm	RF	0.885

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

13	[4], 2020	Using Machine Learning and Electronic Health Records to Predict Postpartum Depression	FFS-RF	SVM	0.78
14	[24], 2020	The application of machine learning in depression	Optimizer	RF	0.9655
15	[12], 2019	Prediction of postpartum depression using machine learning techniques from social media text	LIWC tool	MLPs	0.9163
16	[25], 2019	Using electronic health records and machine learning to predict postpartum depression	Feature comparison, then univariate LR analyses	SVM	0.79

Key: - SFS - Sequential feature selection, NLP- natural language processing, MDI- Mean Decrease in Impurity, LR- logistic regression, SHAP-Shapley additive explanations Processing, FFS-RF -random forest-based filter feature selection, MLPs-Multilayer perceptrons

5. ANALYSIS AND DISCUSSION

Analysis of reviewed publications found that a total of 14 feature selection methods and 11 ML algorithms were used in all the 16 publications studied as shown in tables 3 and 4 respectively. The highest percentage (18.75%) of publications used random forest-based feature selection method while each of the other methods had a single frequency in the remaining articles with a percentage of 6.25% as shown in figure 3. On the other hand, SVM was the most used ML algorithm by 18.75% of the articles, followed by RF and Adaptive Boosting Collaboration algorithms at 12.5% each as shown in figure 4. Each of the other algorithms namely extremely randomized trees, Logistic regression with L2 regularization, Text mining, XGBoost, Voting-Ensemble, Gradient tree boosting algorithm, MLPs and Naïve Bayes had a single frequency in the remaining publications with a percentage of 6.25% of the articles. The performance of all the models developed had an AUC of over 0.70 while the best performance achieved had AUC of 0.9655 which was from a combination of optimizer feature selection method and RF ML algorithm which signified the potential of ML techniques in predicting PPD and other medical conditions.

Table 3: Feature selection methods used and their frequency

Sn	Feature selection Method	Number of publications
1	Sequential feature selection	1
2	Natural language processing	1
3	Gini Importance or Mean Decrease in Impurity	1
4	SelectKBest	1
5	Random forest	3
6	J48	1
7	Shapley additive explanations Processing	1
8	Gradient boosted decision tree	1
9	Vectorization	1
10	ReliefF expRank	1
11	Relief	1
12	Optimizer	1
13	LIWC tool	1
14	Feature comparison, then univariate logistic regression (LR) analyses	1
	TOTAL	16

Table 4: ML algorithms used and their frequency

Sn	ML algorithm	Number of publications
1	Logistic regression with L2 regularization	1
2	Text mining	1
3	Extremely randomized trees	1
4	XGBoost	1
5	Voting-Ensemble	1
6	Adaptive Boosting Collaboration	2
7	Gradient tree boosting algorithm	1
8	Naïve Bayes	1
9	Support vector machine	3
10	Random forest	2
11	Multilayer perceptrons (MLPs)	1
	TOTAL	**16**

The high variation in feature selection methods and ML algorithms used could be a revelation that scientists were yet to settle on the best parameters for an optimal prediction model which qualified the need for continued research on the subject matter. The variation could also have arisen from the need to address specific research objectives which were the focus of the different authors. The Naïve Bayes classifier which tied with logistic regression as the best performing algorithm from the related studies maintained its performance (AUC of 0.93) even though it did not perform a well as other models like AdaBoost (AUC of 0.96), RF (0.9655) and Adaptive Boosting Collaboration (AUC of 0.94). The consistency could not be explained since the feature selection method used from related studies was not reported.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

RF, SVM and Adaptive Boosting Collaboration algorithms drew more interest among the researchers and produced a mixed performance when combined with different feature selection methods. RF algorithm which produced the best model (AUC of 0.9655) when used with optimizer feature selection method also produced the second lowest performance (AUC of 0.75) when combined with ReliefF expRank feature selection method. SVM ML algorithm which had the highest frequency of use by the different authors produced a low performance in all the instances (AUC <0.80) when used with FFS-RF and feature comparison followed by univariate LR analyses feature selection methods. The mixed performance by the algorithm could be an indicator that the high performing combination needed finetuning with repeated trials in different environment to confirm their reliability while as other studies are undertaken to explain the cause of the low performing combinations. The combination of SVM ML algorithm and FFS-RF feature selection methods produced a lower performance outcome (AUC of 0.78) which was the same in two different studies. More studies are needed understand the cause of the performance which should also be done in different environments to confirm the consistency. The lowest performance was from a combination of XGBoost ML algorithm and gradient-boosted decision tree feature selection method (AUC of 0.712). Trials of the same combination of feature selection method and ML algorithm should be undertaken in different environment to evaluate the performance and determine their reliability.

Even though random forest-based feature selection method had the highest frequency of publications, the studies which used the method did not produce the best performing model. Equally, the models created from SVM which was the most frequently used ML algorithm did not produce the most accurate result[4]. The best model was developed from a combination of optimizer feature selection method and RF ML algorithm (AUC of 0.9655) which was almost similar to the next performing model produced from a combination of SelectKBest feature selection method and AdaBoost ML algorithm (AUC=0.96). RF algorithm which produced the best model with Optimizer feature selection method was also used by most studies proving it to be a reliable choice for future models. Whereas these trials provided informative outcomes based on the choices made and their combinations and the environment in which they were used, it was apparent that more research was necessary to finetune the models before adoption. The RF-based feature selection method which was the most popular procedure did not produce the best model when used with RF ML algorithm; its combination with SVM produced a better model[4]. Optimizer feature selection had the least trials but produced the best model. Its high performance needs to be qualified through trials in different environments in combination with other algorithms for comparison. The percentage of publications used for feature selection methods and ML algorithms is shown in figures 3 and 4.

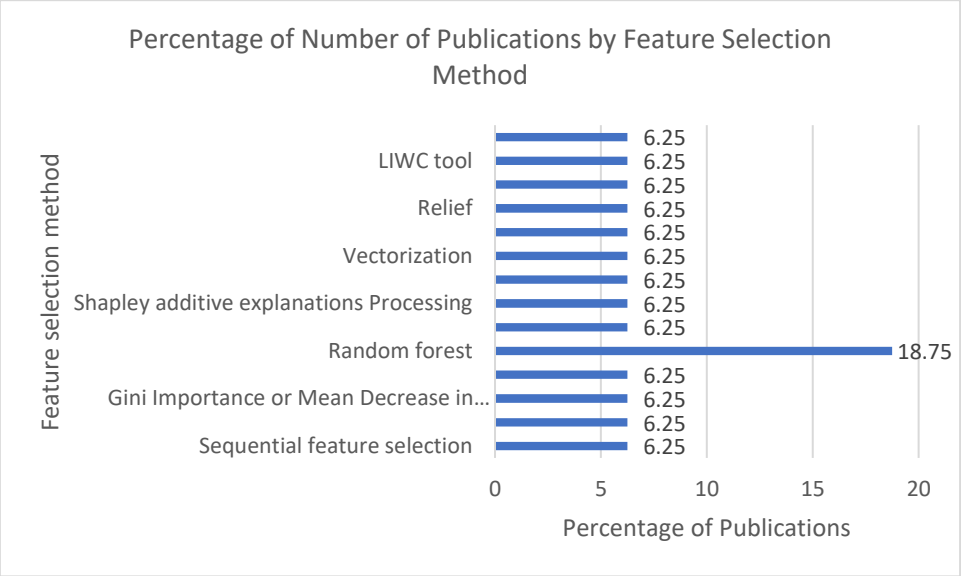


Figure 3: Analysis of feature selection methods by the percentage of publications

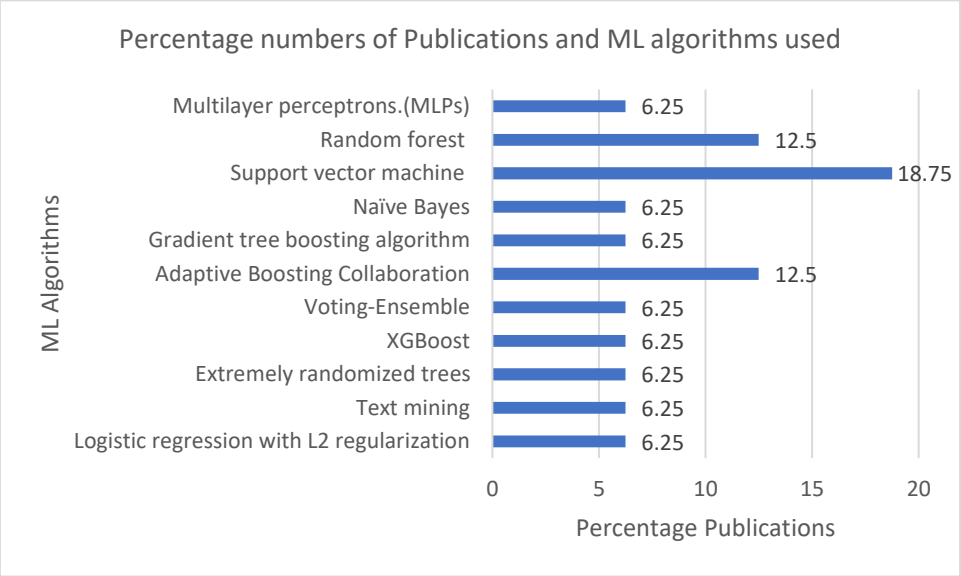


Figure 4: Analysis ML Algorithms by percentage of publications

6. CONCLUSION

The models developed from the review studies produced high performance measures which reflected that ML had great potential to be applied in the prediction of PPD and other medical conditions. Despite the high performance noticed, there was no consistency in the choice and performance of the algorithms used for feature selection and modeling. More trials are required considering the fact that the approaches which were not the most popular choices produced better performance which may not have been expected. The feature selection methods and ML algorithms should be tested under different environments while at the same time interchanging their combinations to compare their performance. It can be concluded that ML algorithms and feature selection methods have not been given enough trials to support a credible analysis of their performance and eventual ranking. More collaborative studies which should consider other contributing parameters in modelling should be carried out for comparison.

REFERENCES

[1] B. Mahesh, *Machine Learning Algorithms -A Review*. 2019. doi: 10.21275/ART20203995.

[2] J.-P. Lai, Y.-M. Chang, C.-H. Chen, and P.-F. Pai, “A survey of machine learning models in renewable energy predictions,” *Appl. Sci.*, vol. 10, no. 17, p. 5975, 2020.

[3] Y. Zhang, S. Wang, A. Hermann, R. Joly, and J. Pathak, “Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women,” *J. Affect. Disord.*, vol. 279, pp. 1–8, Jan. 2021, doi: 10.1016/j.jad.2020.09.113.

[4] W. Zhang, H. Liu, V. M. B. Silenzio, P. Qiu, and W. Gong, “Machine learning models for the prediction of postpartum depression: application and comparison based on a cohort study,” *JMIR Med. Inform.*, vol. 8, no. 4, p. e15516, 2020.

[5] G. Hackeling, *Mastering machine learning with scikit-learn: apply effective learning algorithms to real-world problems using scikit-learn*. Birmingham: Packt Publ, 2014.

[6] Y. E. Alharahsheh and M. A. Abdullah, “Predicting Individuals Mental Health Status in Kenya using Machine Learning Methods,” in *2021 12th International Conference on Information and Communication Systems (ICICS)*, 2021, pp. 94–98.

[7] J. DONG *et al.*, “The application of machine learning in depression,” *Adv. Psychol. Sci.*, vol. 28, no. 2, p. 266, 2020.

[8] M. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, “An in-depth analysis of machine learning approaches to predict depression,” *Curr. Res. Behav. Sci.*, vol. 2, p. 100044, 2021.

[9] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: an overview,” in *Journal of physics: conference series*, 2018, vol. 1142, no. 1, p. 012012.

[10] S. Kumar, N. Kumar, and S. Vivekadhish, “Millennium development goals (MDGS) to sustainable development goals (SDGS): Addressing unfinished agenda and strengthening sustainable development and partnership,” *Indian J. Community Med. Off. Publ. Indian Assoc. Prev. Soc. Med.*, vol. 41, no. 1, p. 1, 2016.

[11] J. Hahn-Holbrook, T. Cornwell-Hinrichs, and I. Anaya, “Economic and health predictors of national postpartum depression prevalence: a systematic review, meta-analysis, and meta-regression of 291 studies from 56 countries,” *Front. Psychiatry*, vol. 8, p. 248, 2018.

[12] I. Fatima, B. U. D. Abbasi, S. Khan, M. Al-Saeed, H. F. Ahmad, and R. Mumtaz, “Prediction of postpartum depression using machine learning techniques from social media text,” *Expert Syst.*, vol. 36, no. 4, p. e12409, 2019, doi: 10.1111/exsy.12409.

[13] K. Saqib, A. F. Khan, and Z. A. Butt, "Machine Learning Methods for Predicting Postpartum Depression: Scoping Review," *JMIR Ment. Health*, vol. 8, no. 11, p. e29838, 2021.

[14] P. Cellini, A. Pigoni, G. Delvecchio, C. Moltrasio, and P. Brambilla, "Machine learning in the prediction of postpartum depression: A review," *J. Affect. Disord.*, Apr. 2022, doi: 10.1016/j.jad.2022.04.093.

[15] M. Usman, S. Haris, and A. C. M. Fong, "Prediction of Depression using Machine Learning Techniques: A Review of Existing Literature," in *2020 IEEE 2nd International Workshop on System Biology and Biomedical Systems (SBBS)*, Dec. 2020, pp. 1–3. doi: 10.1109/SBBS50483.2020.9314940.

[16] M. B. Carneiro, M. W. Moreira, S. S. Pereira, E. L. Gallindo, and J. J. Rodrigues, "Recommender System for Postpartum Depression Monitoring based on Sentiment Analysis," in *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*, 2021, pp. 1–6.

[17] S. Andersson, D. R. Bathula, S. I. Iliadis, M. Walter, and A. Skalkidou, "Predicting women with depressive symptoms postpartum with machine learning methods," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, 2021.

[18] P. Mazumder and S. Baruah, "A Community Based Study for Early Detection of Postpartum Depression using Improved Data Mining Techniques," in *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2021, pp. 1–7.

[19] G. Amit *et al.*, "Estimation of postpartum depression risk from electronic health records using machine learning," *BMC Pregnancy Childbirth*, vol. 21, no. 1, pp. 1–10, 2021.

[20] E. Hochman *et al.*, "Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study," *Depress. Anxiety*, vol. 38, no. 4, pp. 400–411, 2021.

[21] "JMIR Medical Informatics - Machine Learning Models for the Prediction of Postpartum Depression: Application and Comparison Based on a Cohort Study." https://medinform.jmir.org/2020/4/e15516/?utm_source=TrendMD&utm_medium=cpc&utm_campaign=JMIR_TrendMD_1 (accessed Feb. 25, 2021).

[22] K. PY, R. Dube, S. Barbade, G. Kulkarni, N. Konda, and M. Konkati, "Depression Detection using Machine Learning," *Available SSRN 3851975*, 2021.

[23] E. Valavani *et al.*, "Data-Driven Insights towards Risk Assessment of Postpartum Depression.," in *BIOSIGNALS*, 2020, pp. 382–389.

[24] D. Shin, K. J. Lee, T. Adeluwa, and J. Hur, "Machine learning-based predictive modeling of postpartum depression," *J. Clin. Med.*, vol. 9, no. 9, p. 2899, 2020.

[25] S. Wang, J. Pathak, and Y. Zhang, "Using Electronic Health Records and Machine Learning to Predict Postpartum Depression.," *Stud. Health Technol. Inform.*, vol. 264, pp. 888–892, 2019.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 213-222, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1006

Multi-Criteria Parameter Factor in the Assessment of Public Cloud Services Providers' Associated Risks

Omojokun Gabriel Aju
Department of Computer Science
Adekunle Ajasin University
Akungba-Akoko, Nigeria

Abstract: The cloud computing technology is indisputably assisting individuals, businesses and institutions in increasing their capability, productivity and efficiency. It has eliminated the border restriction of obtaining specialized computing resources and expertise that are not available locally to the organizations and individuals without the need to invest in new infrastructure. However, in spite of the immeasurable benefits the technology promises, it has also raised major challenges, particularly as regards its information security and service availability, stressing the need for an elaborate technique of assessing the associated risk of selecting a particular public cloud services provider among available alternatives using the right risk criteria. This paper proposes an expanded multi-criteria risk parameters for the evaluation of public cloud service providers' associated risk to produce a more accurate results that meets the dynamic nature of the cloud technology.

Keywords: Cloud Computing, Cloud Service Providers, Risk Criteria, Information Security, Service Availability.

1. INTRODUCTION

The fast pace at which the organizations and individuals are embracing cloud computing services as the new major milestone in computing technology is a reflection that indeed the technology is an inventions breakthrough, as the technology delivers hosted services over the Internet thereby enabling the organizations to increase their capabilities in meeting computing resources demands while avoiding significant investments in physical infrastructure, training, personnel and software licensing.

Various organizations including commercial industries, academic institutions, military and other government agencies are fast embracing the cloud technology because of its promised benefits of cost reduction, efficiency and resources flexibility that provides organizations with the ability to use specialized computing resources without investing in new infrastructure.

In spite of the undisputed benefits that cloud technology brings, concerns are also being expressed as regards the challenges of the technology, including the potential loss of control of the customers' assets (Information Security and Privacy) by the cloud services providers (Yunchuan et al., 2014; Michael et al., 2015; Domingo-Ferrer et al., 2019; Li et al., 2019; Irshad et al., 2021) and the inability of the providers to guarantee constant availability of the public cloud networks (Liliana et al., 2014; Velliangiri et al, 2020; Qureshi et al., 2020). The fact that public cloud services are shared, externally provided and offered over internet network where users are able to gain access to computing resources from anywhere also makes the services more vulnerable to all forms of attacks (Mohiuddin et al., 2019; Abdurachman et al., 2019; Qureshi et al., 2020; Deebak et al., 2020).

The real and perceived concerns of providing, accessing and controlling services in externally provided multi-tenant cloud environments also slow or preclude the migration of services by major prospective organizations to the public cloud (Nautiyal and Wadhwa, 2019; Dong et al., 2019; Xu et al., 2019). Like every other inventions of technologies, there are various risks that are associated with public cloud services environment arising from the data security and privacy, network availability and performance, systems interoperability, governance and compliance complexity, among others.

The risk assessment in public cloud environment is so challenging compare to the traditional computing environment due to the cloud unique characteristics of on-demand self-service, multi-tenancy and rapid elasticity, which makes the technology more complex and dynamic in nature. Providing a service that meets the needs of subscribers is as important as entrusting adequate level of confidence on the users to ensure that they are taking the right decision of embracing such service, and such certainties of decisions can only be affirmed if there is a way to evaluate the risk associated with such decision, as it is expected that before initiating any substantive contract negotiations or operational integration with a cloud service provider, the prospective consumer should evaluate the cloud provider's competency and commitment to deliver the desired services over the target timeframe while meeting the stipulated service availability and security levels.

Therefore, a more risk dimensional focus area and parameter criteria for the accurate assessment of associated risk of the service environment will go a long way in providing some degree of confidence to the cloud service consumers in assisting them to make right selection decision.

2. PUBLIC CLOUD CHALLENGES

The International Organization for Standardization in ISO 31000:2009 defines risk as the effect of uncertainty on objectives (ISO 31000, 2009). It is expressed as a combination of the consequences of an event and the associated probability of occurrence with the potential to

influence the achievement of an organization's objectives (Berg, 2010). Objectives can have different aspects such as financial, political, reputation or environmental goals depending on the individuals or organizations and can apply at different levels like strategic, project, product or process. It is not therefore unexpected that every individual and organization strives to assess the level of risk associated with their choices at every point in time in order to be guided in their decision making.

According to Stoneburner et al (2002), risk assessment is a process of assessing identified risks in term of their potential severity of loss and possibility of occurrence within a given timeframe. It involves three processes, namely, risk identification, risk analysis and risk evaluation. However, Cloud risk assessment is defined as a dynamic, step by step, repeatable process used to produce an understanding of cloud risks associated with relinquishing control of data or management of services to an external service provider (Akinrolabu et al., 2019). Various research studies have identified different concerns as the major fears of the potential public cloud consumers in the process of adopting appropriate public cloud providers. Daniele & Giles (2009), the Cloud Security Alliance (CSA, 2010); Shukla (2014); Mazhar et al (2015); Odun-AYO (2018); Wu et al (2019); Karajeh et al (2020) and Irshad et al (2021) identified information security as the major fear of the cloud services consumers.

Charanya et al (2013); Mohammed et al (2013); Srivastava and Khan (2018); Deebak et al (2020) and Alghofaili et al (2021) considered data sovereignty as a major challenge in the public cloud environment, the researchers based their conclusion on the fact that most public cloud services providers or their data centres are mostly located outside the jurisdiction of the service consumers, or when such service providers or their data centres are located within the jurisdiction of the service consumers, they are mostly owned by third party agents. Sah et al (2014); Siddiqui (2019); Dong et al (2019); Abdurachman et al (2019); and Tabrizchi and Rafsanjani (2020) insisted that the multi-tenancy method of service delivery using resources pooling through the virtualization technology creates great security risks, as public cloud service providers deliver services to multiple customers (tenants) by sharing the same computing resources.

Eric et al (2012); Hashizume et al (2013); Kamal et al (2014); Yunchuan et al (2014); Michael et al (2015); Paul et al (2018); Kumar et al (2018); Verma and Sharma (2019); Wu et al (2019) and Alghofaili et al (2021) reported data security and privacy as the major obstacles hampering the widespread adoption of public cloud computing. The studies observed that most of the services (SaaS and PaaS) providers do not have access to the physical security system of data centres, they mostly rely on third party to achieve full data security and the fact that consumers are to handover their data to a third party is a major challenge.

Srivastava and Khan (2018); Verma and Sharma (2019); Dong et al (2019) reported network availability, performance unpredictability and system interoperability as the main challenges facing the organizations' decision of moving to the public cloud. Liliana et al (2014); Singh (2017) and Alghofaili et al (2021) presented lack of interoperability standards as a threat within the public cloud environment since there is neither standardized communication between and within public cloud providers nor standardized data export format. It is therefore difficult to migrate from one cloud service provider to another or bring back data and process it in-house, this makes it difficult to establish security frameworks for cloud heterogeneous environments.

There are few risk assessment frameworks and models which are designed for the public cloud consumers to assist them in their selection of public cloud services providers during the cloud adoption based on various criteria and risk focus using different methodologies, such as Chandran and Angepat (2010), QUIRC by Prasad and Ben (2010), SecAgreement by Matthew and Rose (2012) and Microsoft by Greg and Pierre (2016), among others. However, this study is exclusively concern with the suitability of the existing parametric criteria and the focus area of risk in the public clouds to derive accurate risk values for the purpose of decision making in the selection of the appropriate public cloud provider among many alternatives.

3. THE EXISTING RISK CRITERIA

The issue of designing risks assessment frameworks and models for cloud environment started in 2009 with the design and publication of Cloud Computing Information Assurance Framework by the European Union Agency for Network and Information Security (ENISA) (Daniele and Giles, 2009). The framework followed ISO/IEC 27005:2008 risk level estimation approach for the traditional information systems and categorized the public clouds security risks into four groups: policy and organization risks, technical risks, legal risks and the other scenarios not specific to cloud technology. The framework uses generic qualitative approach while focusing on the information security within the cloud environment.

Prasad and Ben (2010) presented a quantitative risk assessment framework (QUIRC) for public cloud security based on the Federal Information Processing Standards (FIPS) of the US Federal Information Security Management Act (FISMA) and adapted the Wide-band Delphi method of rankings which is based on experts opinion about the likelihood and consequence of threats to assess the security risks associated with public cloud services providers. However, the framework was a localized work as it was based on the US Federal Information Security Management Act (FISMA) for information processing within the public sector in the United States of America (USA) which makes it a public policy.

Chandran and Angepat (2010) proposed a public cloud risk assessment framework based on Trust Matrix Approach for security risk analysis to ensure that formal risk assessments are aligned with the enterprise-wide framework to facilitate transparency and increase trust level between the cloud customers and the cloud providers. The framework used two variables, namely "Data Cost" and "Providers' History" as risk criteria. In "data cost" users can assign a cost to data based on the data's criticality whereas "Provider's history" includes the record of the past services provided by the provider to consumers. The framework also focuses on the information security aspect of the public cloud technology.

Peiyu and Dong (2011) produced a cloud information security risk assessment model for public cloud services consumers based on theory of Analytic Hierarchy Process (AHP) by listing eight (8) kinds of threats to cloud security principles and their corresponding correlation coefficients to get the information security risk assessment of public cloud service

provider. The model specifically focus on information security aspect of the public cloud using security as the only criteria and eight security threats that are peculiar to cloud technology as sub-criteria.

Feng et al (2012) presented a risk management framework for public cloud consumers on the basis of service providers' previous works focusing on the public cloud information security. The aim of the work was to assist the public cloud consumers to ascertain the risk associated with adopting a particular public cloud service provider by reviewing the service providers' previous services to their customers. The framework analyses the security status of cloud service providers by reviewing historical incidents associated with the service providers and introduces the involvement of third party assessment agency to ensure thorough analysis of the provider's capability. Matthew and Rose (2012) developed a cloud risk assessment model (SecAgreement) based on the Service Level Agreement (SLA) negotiation standard to allow security measures to be expressed on service description terms and service level agreement. The approach defines a cloud service matchmaking algorithm to assess and rank the SLA by their risk, allowing the consumers to quantify risk, identify any policy compliance gaps that might exist, and thus select the public cloud services providers that best meet their security needs.

Mouna et al (2015) proposed a multidimensional approach towards a quantitative assessment of information security risks in the public clouds model. The model illustrates how a quantitative risk assessment of public cloud service providers can be carried out based on a systematic, extendable and modular approach. The model views information security risks as segmentation of the public cloud world according to its dimensions, where a dimension can be defined as an elementary aspect of risk sphere. The model uses a new approach to threats classification using dimensional method and a quantitative assessment of the associated risks based on the number of identified dimensions allowing the model to be modular and extendable in nature, although, no specific risk criteria was stated for the assessment purpose.

Greg and Pierre (2016) designed a cloud risk decision framework that was based on the ISO 31000 standard (ISO 31000, 2009) to assist the cloud consumers to take appropriate risk decision before moving to the cloud by using the framework as a template in assessing the risks associated with a particular cloud providers. The framework based the cost of the data security breach in the public clouds to the prospective consumers into four groups of operational risk, market and finance risk, strategies risks and compliance risks and used qualitative assessment approach to evaluate the risk level based on these four risk types. Cayirci et al (2016) presented a public cloud adoption risk assessment model (CARAM) based on the three existing frameworks of ENISA, Cloud Security Alliance's Consensus Assessment Initiative Questionnaire (CAIQ) and the French National Commission on Informatics and Liberty (CNIL) developed in Europe for assisting public cloud services consumer to select a cloud services provider that fits their security risk profile best. The model essentially focuses on information security risk by adopting the ENISA's thirty-five (35) security risk elements and the Cloud Security Alliance's CAIQ eleven (11) security risk control areas.

Sivasubramanian et al (2017) produced a cloud risk assessment model for public cloud services consumers based on the probability of an incident occurring, which is mapped against the estimated negative impact. The model used the Information Assets and Risk, Privacy and Confidentiality Concerns, Data Governance for its risk assessment which is principally based of the Data Cost variable. The Expression of Needs and Identification of Security Objectives (EBIOS) method for evaluating and treating risks, that aims to determine the security actions to implement which focuses on six key categories of security objectives (SO) (i.e. Confidentiality, integrity, availability, multi- party trust, mutual audit ability and usability) and the Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) method of assessment of vulnerabilities and threats on the basis of the operating assets of a company were adopted in the assessment. The authors have followed the traditional route to security risk assessment, concentrating on the local organization, their critical assets (data), threats, and likelihood of impact, without paying attention to the supplier network nor fully understanding its interrelated consequences.

Akinrolabu et al (2019) presented a quantitative risk assessment model, termed Cyber Supply Chain Cloud Risk Assessment (CSCCRA) based on the systematic analysis of cloud risks, the visual representation of the cloud supply chain, and the assessment of the cyber security posture of cloud suppliers through the use of experts and evaluation of supply chain. The authors adopt the use of information of the parties involved in the development, hosting, management, monitoring and use of the cloud services (i.e. the supply chain).

The model takes a multi-disciplinary approach to assessing the dynamic, evolving and interconnected risks in the cloud, applying different knowledge areas in the identification, analysis, and evaluation of these risks. It combines factors such as security, supplier selection, systems thinking, decision support systems, quantitative risk modelling, and supply chain mapping in a multi-stage approach.

It is obvious that the existing frameworks and models cannot provide accurate risks level evaluation of the public cloud services providers to assist the prospective consumers in their selection decision making processes (Alghofaili et al (2021), this is because the existing models and frameworks focuses the assessment of associated risks of public cloud providers only on the information security aspect of the public cloud services and therefore limit their risk criteria on information security. The motivation for this study is therefore drawn from the limitations of the existing risk focus areas and parametric criteria for the purpose of evaluating the associated risks of specified public cloud providers for the purpose of selecting the best alternative among the various service providers.

4. PROPOSED RISK CRITERIA

It has been observed from the reviewed literatures that the existing models focus on the security of information in the cloud in their risk assessment process as a means of selecting the appropriate public cloud provider. However, the rapid expansion of public clouds services from the predominantly established data storage services facilities to many other services, such as real-time enterprise networking, educational mobility solutions, digital videos services, financial and industrial intensive processing solutions, national satellite monitoring services, offices on motion services, among other numerous services have made the services availability and performance a major area of concern (Yogeshwaran et al., 2017; Dong et al., 2019; Jouini and Rabai, 2019).

Unfortunately, none of the existing public cloud providers' risk assessment work in the reviewed literature has considered the public cloud services availability and performance as a major focus in the assessment of risks for the purpose of selecting appropriate public cloud service provider, resulting in the restriction of the criteria for the risk assessment to information security area. To address this deficiency in the risk focus areas and the selection of the risk criteria, this study is hereby proposing wider areas of risk focus and assessment criteria.

4.1 The Risk Area of Focus

4.1.1 Information Security

The existing public cloud service providers risk assessment models, such as Prasad and Ben (2010); Chandran and Angepat (2010); Peiyu and Dong (2011); Matthew and Rose (2012); Mouna et al (2015); Greg and Pierre (2016); Cayirci et al (2016); Sivasubramanian et al (2017) and Akinrolabu et al (2019) all focused on the information security as an area of risk within the public cloud environment as reflected in the choice of risk criteria used in the models. As cloud services are provided through the internet technology, it causes the cloud computing systems to inherits those security challenges that are peculiar to the internet technology resulting to number of vulnerabilities within the public cloud services environment as data is being indiscriminately shared among the varied systems which affects the validity, quality and security of the data in the public clouds (Rana and Mohammed, 2016; Qureshi et al., 2020 and Deebak et al., 2020). Therefore, taking the security of data (Assets) to be placed in the cloud as a factor in the process of assessing the associated risks of cloud service providers to examine the level of protection by the service provider(s) cannot be overemphasised.

4.1.2 Service Availability and Performance

Liliana et al (2014); Srivastava and Khan (2018); Verma and Sharma (2019); Dong et al (2019); Aldribi et al (2020) and Alghofaili et al (2021) presented cloud network availability, performance unpredictability and system interoperability as major challenges facing organizations' decision of moving to the cloud. The authors posited that the availability and performance of public cloud services are heavily dependent on the supporting technological infrastructure, and that the available bandwidth, reliability and resiliency of local and international network connections could have a significant impact on consumers' public cloud experience. Yogeshwaran et al (2017); Mohiuddin et al (2019); Nautiyal and Wadhwa (2019) and Velliangiri et al (2020) observed that the expansion of public cloud services beyond the data storage services to real-time enterprise networking, education mobility solutions, digital videos services, financial and industrial intensive processing solutions has introduced cloud service availability and performance as a serious risk concern to the consumers and a major factor in the adoption of public clouds.

A research from the University of California tracked the availability and outages of four major cloud providers in the United States of America and found out that overloads on the cloud systems caused programming errors resulting in system crashes and failures. Likewise, due to inefficient business continuity and backup recovery mechanism, public cloud services experience periods of unavailability ranging from minutes to days, resulting in loss of confidence among the customers which brought up fresh debates on the capability of the cloud technology in handling certain critical computing services. For example, In March 2018, Amazon Web Services (AWS) was hit by a cloud outage that silenced Amazon's Alexa and affected hundreds of enterprise services including Atlassian, Slack, and Twilio. The outage happened in the data centres in Virginia when the Direct Connect dedicated links from AWS North Virginia region to other server warehouses and premises on the East Coast got disabled and the outage lasted for about 4 days.

More worrisome, natural disasters also present significant risks in the cloud services environment. For example, in August 2018, Microsoft suffered an outage caused by a severe lightning storm in the San Antonio; Azure's South-Central United States data centre region was down for quite a while. Customers across the world using Active Directory and Visual Studio Team Services faced trouble for more than 24 hours. Therefore, service availability and performance plays a major role in cloud computing as the needs of the customers should be attended to at all times. Regrettably, the existing models practically focused only on information security as an area of risk at the exclusion of services availability and performance.

4.2 The Risk Assessment Criteria

Chandran and Angepat (2010) used "Data Cost" and "Cloud Providers' Service History" as risk criteria to assess the risks associated with the public cloud service providers. Feng et al. (2012) used the "Cloud Providers History" as a risk criteria in the assessment of the public cloud service providers' risks. Matthew and Rose (2012) relied on the information security contractual obligations embedded in Service Level Agreements (SLA) to assess the potential risk associated with a service provider while Daniele and Giles (2009), Mouna et al. (2015), Cayirci et al. (2016); Greg & Pierre (2016); Sivasubramanian (2017) and Akinrolabu et al (2019) used the lists of information security threats within the public clouds to assess the associated risks of the public cloud service providers. These criteria are not sufficient in providing an accurate evaluation of risks level to the public cloud consumers as they are specifically based on the information security risk aspect of the public clouds. While these two risk criteria can be considered as major criteria in the determination of information security risk within the public clouds, they are not able to specifically affect the public cloud services availability and performance; neither do they cover all the public cloud services information security loopholes.

Therefore, in addition to the "Data Cost"(Asset Cost) and "Providers' History" from the existing models (Chandran and Angepat, 2010; Feng et al., 2012; Mouna et al., 2015; Greg and Pierre, 2016) and others, three additional risk criteria of **Service Location, Adopted Technology** and **People** are introduced in this research study as these three criteria have been identified as major cloud risk determinants in the public clouds that have direct and indirect effects on the information security as well as the services availability and performance of the public cloud services.

4.2.1 Data Cost

In assessing the risk that is associated with selecting a particular public cloud provider, it is essential that the value, critically and sensitivity of the data or assets to be transferred to the cloud is recognized as well as the service providers' reputation (Armbrust et al, 2010; Aissaoui et al., 2017; Domingo-Ferrer et al., 2019). The under-classification of data or assets could result in such assets being placed in an inappropriate cloud service that cannot provide expected level

of protection and services. Conversely, over-classification of assets could lead to unnecessary demand of protection and services being specified leading to excessive costs resulting in suitable cloud services providers being rejected (Scott et al, 2010). Therefore it is crucial that the consumers accurately assess the value, criticality and sensitivity of the assets to be placed on the public cloud and correctly classifies it to ensure that the appropriate cloud service provider that meets expected services and protection is shortlisted and selected.

In the process of assessing the assets cost, certain important considerations must be noted as regards the assets and the reputation of the service providers, such as:

i. The owner(s) of the assets/data
ii. The users of the assets/data
iii. The businesses or services supported by the assets/data
iv. The legislation that applies to the information
v. The share values of the service providers and the assets owner(s)
vi. The impact of the assets on the owner's organization and business

Data cost is considered as one of the criteria variables because the consumers can assign a cost to the data based on the data's value, criticality and sensitivity, with its impact on the consumers' organization reputation and service.

4.2.2 Providers' History

Provider's History is considered as another parameter in the process of determining the risk that is associated with a particular public cloud service provider as it includes the record of the past services provided by the service providers to their customers, enabling the consumers to assess the provider(s) service reliability (Chandran & Angepat, 2010; Feng et al (2012). By examining the history of service providers, the prospective consumer would be empowered with information regarding the providers' years of service experience, the nature of rendering services and the industries of the past and existing customers with the customers' locations, the service availability rates of the past services, among other information.

More importantly, the information about the service providers can also reveal the percentage of services directly supply by the service providers and the percentage that is contracted to other parties (subcontractors) and the locations of these subcontractors. The service providers' past and existing relationships with these others parties and the reliability of services of the parties can be examined, this will allow the prospective consumers to rate the service providers.

4.2.3 Service Location

Charanya et al (2013); Mohammed et al (2013); Srivastava and Khan (2018) and Deebak et al (2020) identified data sovereignty risk as a major challenge in the public cloud environment, the researchers based their conclusion on the fact that most public cloud services providers or their data centres are mostly located outside the jurisdiction of the service consumers. The information and data laws differ from country to country; therefore the laws that influence the access of information held by the service providers vary from country to country based on the location of such information (data) or service providers. Hashizume et al (2013); Velliangiri et al (2020) and Alghofaili et al (2021) reported that the movement of data into the public cloud and potentially across and between legal jurisdictions including offshoring of data processing allows certain practices that provide intruders with gates to the information in the cloud, more so that it is difficult to guarantee that a copy of data or its backups are not stored or processed in a certain jurisdiction.

In certain instances, a service provider may be compelled by a foreign law enforcement agency or legally constituted court to provide data belonging to their customers, while legally prohibited from notifying the customer(s) of such disclosure request. In some circumstances, service providers outsource or sub-contract part of the delivery of the service to a third-party leading to additional data sovereignty risks. For example, in August 2014, Microsoft was ordered by a United States Federal Court to turn over customers' data stored in its Republic of Ireland data centre, the Federal Court Judge (Loretta Preska) rejected Microsoft's argument that a United States' search warrant does not extend beyond the country's border (Jaikumar, 2014). Therefore, it is very important for the service consumers to identify the legal jurisdictions in which their data will be stored, processed or transmitted and how the laws of those countries could impact on the confidentiality, integrity, availability and privacy of the data.

Furthermore, certain locations are known for experiencing frequent natural disasters such as flood, earthquake, hurricanes, tsunamis or volcanic eruptions which can affect the cloud service availability and invariably the information availability to the consumers. In 2012, the Atlantic hurricane season saw the arrival of Sandy, the 2nd-costliest hurricane in U.S. history. The floods and power outages wreaked havoc on data centres in New York, New Jersey, Florida and the surrounding areas resulting in the disruption of access to data stored in the public cloud globally, particularly the credit card services for days. The incident perhaps opened up broader discussions around the impact of natural disasters on businesses and services continuity, that nothing is immune to the wrath of Mother Nature, not even the cloud (Uri, 2013).

4.2.4 Adopted Technology

Srivastava and Khan (2018); Verma and Sharma (2019) and Dong et al (2019) reported service availability, performance unpredictability and system interoperability as challenges confronting the organizations' decision of moving to the public cloud. Liliana *et al*. (2014); Srivastava and Khan (2018) and Verma and Sharma (2019) presented the lack of interoperability standards as a major threat within the public cloud environment. Sah et al. (2014); Dong et al., (2019); Mohiuddin et al., (2019); Aldribi et al (2020) and Yang et al (2020) insisted that the multi-tenancy method of service delivery using resource pooling through the virtualization technology creates great privacy and service availability risks, as service providers deliver services to multiple customers (tenants) by sharing the same computing resources.

While resource pooling and sharing has its benefits in terms of costs, it does introduce some form of risks related to either infrastructure virtualization or data commingling that must be considered by the cloud service consumers. Virtualization is an important technology in the delivery of public cloud services as it enables information systems to be abstracted from the underlying hardware using a hypervisor (that is, software that enables a host server to run multiple guest operating systems concurrently).

The most often cited area of concern of this technology is that a malicious customer could exploit vulnerability within the

hypervisor to gain access to another customers' information by performing a 'guest-to-host' or 'guest-to-guest' attack. Also, some cloud services such as SaaS and PaaS use logical controls within the application or platform and supporting infrastructure to isolate access to each customer's data. However, the data are usually commingled within the application, database and back-up systems. This places complete reliance on the quality of the design, implementation and enforcement of access controls within the platforms and applications.

More importantly, denial of service (DoS) attacks is an inherent risk for all Internet facing services. The use of cloud services may increase the risk of such an attack as the aggregation of multiple customers into a single service may present a more attractive target for attackers. Therefore, a customer may suffer associated or collateral damage in form of service unavailability in an attack against a service provider or a co-customer. The service providers adopted protocols and technologies, such as, Anycast, Application Delivery Networks and Content Delivery Networks in distributing network traffics and computer processing can determine the extent of such attack against their services platforms.

These concepts explain the importance of service availability, performance unpredictability and system interoperability as decisive elements in the provision of public cloud services. The elements are product of the technologies being adopted by the cloud providers, such as operating systems, virtualization systems, network systems, cooling system, security protocols (such as, encryption protocols and authentication methods), application programming interface, and database management systems. Unfortunately, these important elements were not considered in the existing works. It is therefore certain that the types of technologies adopted and the adoption rate of new technologies by the public clouds providers play major role in the service delivery of the offered services and such important element cannot be ignored in the process of assessing the risks associated with the public cloud services providers for the purpose of selecting the appropriate provider that meets the requirements of the consumer among the available alternatives.

4.2.5 The People

The peoples' dimension of the public cloud technology which encompasses all its classes of users and their roles cannot be overemphasized in the process of assessing risks associated with public cloud service providers. Gouda et al. (2014); Yunchuan et al. (2014); Michael et al. (2015); Domingo-Ferrer et al., 2019 and Tabrizchi and Rafsanjani (2020 observed that the inability of the public cloud consumers to ascertain the service providers' employees' reliability and trustworthiness, and whether a service provider has appropriate procedures in place to ensure that its personnel are reliable and trustworthy is a common concern for organizations planning to use public cloud services.

Farhad and Sajjad (2012); Kumar et al (2018); Jouini and Rabai (2019); Yang et al (2020) considered public cloud service providers insider threat as a major concern to the public cloud consumers, as unauthorized access to sensitive information by the service provider's employees is a common concern for organizations' planning to use cloud services.

The idea of handing over important data to another company which reliability and trustworthiness of its employees cannot be ascertained worries some individuals and organizations. For example, on 28 February, 2017, an Amazon Web Services Engineer trying to debug a Service Storage System (S3) at their Virginia data centre accidentally typed an incorrect command and much of the Internet including many enterprise platforms and cloud servers critical-mission services were down for 5 hours resulting to cloud services international outage. The outage from the provider that owns roughly a third of the global public cloud market reignited debate on the risks of public cloud (Joseph, 2017).

The public cloud service customers should ascertain the experience and expertise of the key employees and whether the service providers have appropriate procedures in place to make sure their personnel are reliable, trustworthy and do not pose a security risk to their clients. Though, the level of assurance available to the consumers vary significantly depending on the physical location of the service provider's services and its employees, as it may be very easy to ascertain such security check of the service provider's employees if the prospective customer is within the same geographical jurisdiction as the service provider. However, where a service is delivered or supported from another geographical jurisdiction (country) these security checks procedures may be very difficult to undertake or even impossible. In such circumstances, the prospective customers may consider whether the available alternatives to the service provider can provide an equivalent level of assurance. Although, while vetting may prevent service providers from employing someone that has a history of being untrustworthy, it does have its limitations, as vetting that reveals a criminal record may result in a potential employee being rejected. In the same manner, candidates that are untrustworthy but have never been caught or have not been convicted may not be identified. So also, previously trustworthy employees may become untrustworthy if they become disgruntled or their personal circumstances change.

Interestingly, the on-demand self-service characteristic of cloud computing also introduces security concerns as the customers' registration processes (usually, web-based self-registration) are not always robust to confirm a customer's identity. This weakness can allow a malicious customer(s) to register for services to be used for malicious activities that may include attempting to subvert the access controls to gain unauthorized access to another customer's data. These human involvements in the activities of the cloud services have major implications on both the service providers and consumers organizations, and this should be a major criterion for consideration in the assessment of the risks associated with public cloud service providers.

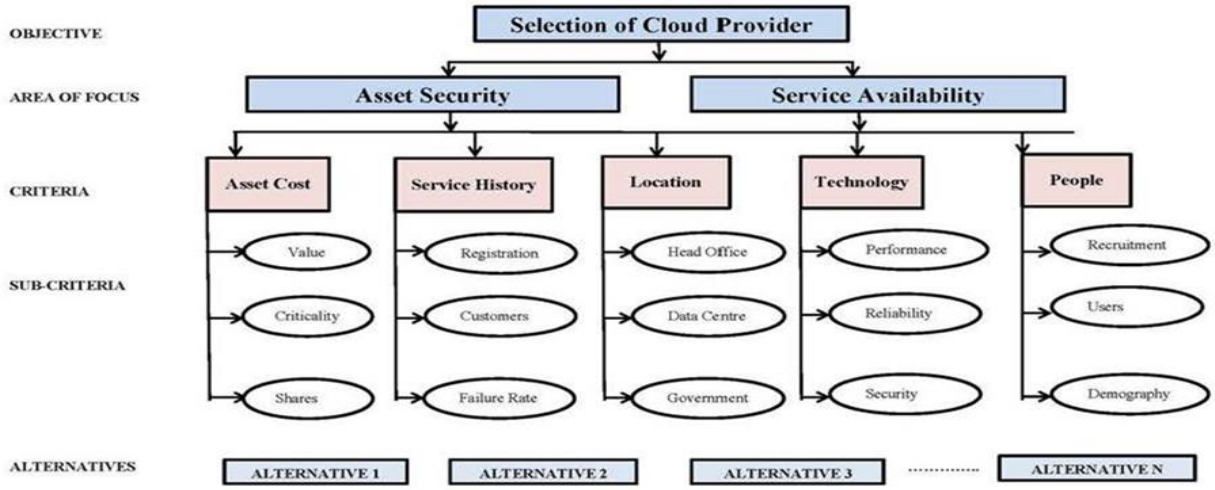


Figure 1: The Hierarchy of the Selection Parameters and the service Providers Alternatives

5. CONCLUSION

The challenges faced by the public cloud consumers in the process of selecting appropriate public cloud providers that meet their organization's requirements. The selection decision becomes more complicated in case of multiple service providers, conflicting criteria and imprecise parameters, stressing the need for comprehensive criteria that factors the dynamic nature of public clouds computing systems. Therefore, in addition to the two major parametric criteria (Data Cost and Provider's History) adopted by the majority of the existing models and frameworks, newly established risk criteria have been added, these are Service Location, Adopted Technology and the People.

The extension of the risk criteria as assessment parameters is made necessary as a result of the extension of risk assessment focus areas to include service availability and performance in addition to the information security, so as to produce a more accurate risk assessment of the public cloud service providers that can assist the public cloud consumers to make appropriate selection decision among the available public cloud services providers.

6. REFERENCES

[1] Abdurachman, E.; Gaol, F.L and Soewito, B. (2019). Survey on Threats and Risks in the Cloud Computing Environment. Procedia Computer Science, Vol.161, pp. 1325–1332.

[2] Aissaoui, K.; Idar, H. A.; Belhadaoui, H and Rifi, M. (2017). Survey on Data Remanence in Cloud Computing Environment. In the Proceedings of the International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS, 2017), Fez, Morocco, 19–20 April 2017.

[3] Akinrolabu, O., Nurse, J. R. C., Martin, A and New, S. (2019). Cyber risk assessment in cloud provider Computers and security Vol. 87: 101600.

[4] Aldribi, A.; Traoré, I.; Moa, B and Nwamuo, O. (2020). Hypervisor-based cloud intrusion detection through online multivariate statistical change tracking. Computers and Security. Vol. 88: 101646.

[5] Alghofaili, Y., Albattah, A., Alrajeh, N., Rassam, M. A and Al-rimy, B. A. S. (2021). Secure Cloud Infrastructure: A Survey on Issues, Current Solutions, and Open Challenges. Applied. Science 2021, 11, 9005. https://doi.org/10.3390/app11199005.

[6] Armbrust, M., Fox, A., Katz, R., Konwinski, A., et al. 2010. A View of Cloud Computing. Communication of the ACM, Vol. 53, No. 4, pp. 50-58.

[7] Berg, H. (2010). Risk Management: Procedures, Methods and Experiences, Bundesamt für Strahlenschutz, Salzgitter, Germany, 2010.

[8] Cayirci, E., Garaga, A., Santana de Oliveira, A., Roudier, Y. (2016). A Risk Assessment Model for Selecting Cloud Service Providers. Journal of Cloud Computing: Advances, Systems and Applications. Vol. 5, No. 14.

[9] Chandran, S. P. and Angepat, M. (2010). Cloud Computing: Analyzing the risk involved in cloud computing environments. In Proceedings of the International Conference on Natural Sciences and Engineering, Sweden, pp. 2–4

[10] Charanya, R., Aramudhan, M., Mohan, K. and Nithya, S. (2013). Levels of Security Issues in Cloud Computing. International Journal of Engineering and Technology, Vol. 5, No. 2. pp. 1912-1920.

[11] Cloud Security Alliance (CSA). (2010). 'Top Threats to Cloud Computing V1.0. www.cloudsecurityalliance.org/topthreats (Accessed: 13 December, 2021)

[12] Daniele, C. and Giles, H. (2009). Cloud Computing: Benefits, risks and recommendations for information security. ENISA, Crete (Greece). https://www.enisa.europa.eu/publications/cloud-computing-risk-assessment

[13] Deebak, B.; Al-Turjman, F and Mostarda, L. (2020). Seamless secure anonymous authentication for cloud-based mobile edge computing. Journal of Computers and Electrical Engineering. Vol. 87, 106782.

[14] Domingo-Ferrer, J.; Farràs, O.; Ribes-González, J and Sánchez, D. (2019). Privacy-Preserving Cloud Computing on Sensitive Data: A Survey of Methods, Products and Challenges. Computer Communications. Vol.140, pp. 38–60.

[15] Dong, S.; Abbas, K and Jain, R. (2019). A Survey on Distributed Denial of Service (DDoS) Attacks in SDN and Cloud Computing Environments. IEEE Access, Vol. 7, pp. 80813–80828.

[16] Eric, H., Ed, M. and Karen, G. (2012). Risk Management for Cloud Computing. TechTarget, MA 02466, USA.

[17] Farhad, S. G and Sajjad, H. (2012). Security Challenges in Cloud Computing with More Emphasis on Trust and Privacy. International Journal on Scientific and Technology Research. Vol. 1, No.6. pp. 49-54

[18] Feng, X., Yong, P., Wei, Z., et al. (2012). A Risk Management Framework for Cloud Computing. IEEE 2nd International Conference on Cloud Computing and Intelligent Systems, pp. 476-480. China.

[19] Gouda, K. C., Dines, D., Anurag, P., *et al*. (2014). Migration Management in Cloud Computing. International Journal of Engineering Trends and Technology, Vol. 12, No. 9. pp. 466-472

[20] Greg, S. and Pierre, N. (2016). Cloud Risk Decision Framework: Principles and Risk-Based Decision-Making for Cloud-Based Computing. Microsoft Inc., USA.

[21] Hashizume, K., Rosado, D., Medina, E.F., et al. (2013). An analysis of security issues for cloud computing. Journal of Internet Services and Applications. Vol. 4, No. 5

[22] International Organization for Standardisation (ISO). (2009). ISO 31000: Risk management - Principles and guidelines. https://www.iso.org/standard/43170.html

[23] Irshad, A.; Chaudhry, S.A.; Alomari, O.A.; Yahya, K and Kumar, N. (2021). A Novel Pairing-Free Lightweight Authentication Protocol for Mobile Cloud Computing Framework. IEEE System Journal. Vol. 15, pp. 3664–3672.

[24] Jaikumar, V. 2014. Data Security and Privacy Issues in the Cloud: Microsoft ordered to turn over customer data stored in the cloud. [Online]. Available: https://www.computerworld.com/article/2490690/technology-law-regulation/microsoft-ordered-to-turn-over-customer-data-stored-in-the-cloud.html. (Accessed: 21 October, 2021).

[25] Joseph, T. (2017). The 10 Biggest Cloud Outages of 2017. https://www.crn.com/slide-shows/cloud/300089786/the-10-biggest-cloud-outages-of-2017-so-far.htm/pgno/0/5. (Accessed on: 4 September, 2021).

[26] Jouini, M and Rabai, L.B.A. (2019). A Security Framework for Secure Cloud Computing Environments. In Cloud Security: Concepts, Methodologies, Tools, and Applications; IGI Global: Hershey, PA, USA, pp. 249–263.

[27] Kamal, K. H., Ruchi, D. and Rakesh, R. (2014). Security and Privacy Issues of Cloud and Grid Computing Networks. International Journal on Computational Sciences and Applications. Vol. 4, No. 1. pp 83-91.

[28] Karajeh, H.; Maqableh, M and Masa'deh, R. (2020). Privacy and Security Issues of Cloud Computing Environment. In the Proceedings of the 23rd IBIMA Conference Vision, Valencia, Spain, 13–14 May 2020.

[29] Kumar, P. R.; Raj, P. H and Jelciana, P. (2018). Exploring Data Security Issues and Solutions in Cloud Computing. Procedia Computer Science. Vol.125, pp. 691–697.

[30] Li, H., Liu, L., Lan, C., Wang, C and Guo, H. (2020). Lattice-Based Privacy-Preserving and Forward Secure Cloud Storage Public Auditing Scheme. IEEE Access, Vol. 8, pp. 86797-86809.

[31] Liliana, F.B. S., Diogo, A.B. F., Joao, V.G., et al. (2014). Cloud security: state of the art, in: Security,

Privacy and Trust in Cloud Systems. Springer, Berlin, Heidelberg. pp. 3–44.

[32] Matthew, H. L. and Rose, G. (2012). SecAgreement: Advancing Security Risk Calculations in Cloud Services. In proceedings of 2012 IEEE 8th World Congress on Services, Honolulu, HI. pp. 133-140.

[33] Mazhar, A., Samee, U. K. and Athanasios, V. (2015). Security in cloud computing: Opportunities and challenges. Information Science, Elsevier. Vol. 305, pp. 357-383.

[34] Michael, M., Rajiv, R., Lizhe, W., et al. (2015). CloudGenius: a hybrid decision support method for automating the migration of web application clusters to public clouds. IEEE Transaction on Computers, Vol. 64, No. 5. pp. 1336-1348.

[35] Mohammed, A. A., Ben, S. and Eric, P. (2013). A Survey on Data Security Issues in Cloud Computing: From Single to Multi-Clouds. Journal of Software, Vol. 8, No. 5. pp. 1068-1078.

[36] Mohiuddin, I.; Almogren, A.; Alrubaian, M and Al-Qurishi, M. (2019). Analysis of Network Issues and their Impact on Cloud Storage. In the Proceedings of the 2nd International Conference on Computer Applications & Information Security (ICCAIS, 2019), Riyadh, Saudi Arabia, 1–3 May 2019.

[37] Mouna, J., Latifa, B. R. and Ridha, K. (2015). A Multidimensional Approach Towards a Quantitative Assessment of Security Threats. In Proceedings of the 6th International Conference on Ambient Systems, Networks and Technologies. Procedia Computer Science 52, Elsevier, pp. 507-514.

[38] Nautiyal, S and Wadhwa, S. (2019). A Comparative Approach to Mitigate Economic Denial of Sustainability (EDoS) in a Cloud Environment. In the Proceedings of the 4th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 21–22 November 2019.

[39] Odun-Ayo, I., Agono, F and Misra, S. (2018). Cloud Migration: Issues and Developments. In the Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2018), Hong Kong. Vol. 1.

[40] Paul, V., Pandita, S and Randiva, M. (2018). Cloud Computing Review. International Research Journal of Engineering and Technology (IRJET), Vol. 5 Issue 3. pp. 1454-1456.

[41] Peiyu, L. and Dong, L. (2011). The New risk assessment model for information system in Cloud Computing environment, Procedia Engineering 15, Elsevier, pp. 3200 – 3204.

[42] Prasad, S. and Ben, W. (2010). QUIRC: A Quantitative Impact and Risk Assessment Framework for Cloud Security. In Proceedings of the IEEE 3rd International Conference on Cloud Computing, pp. 280-288.

[43] Qureshi, A.; Dashti, W.; Jahangeer, A and Zafar, A. (2020). Security Challenges over Cloud Environment from Service Provider Prospective. Cloud Computing and Data Science, Vol.1, pp.1–48.

[44] Rana, A. and Mohammad, A. (2016). Risk Management Framework for Cloud Computing: A critical Review. International Journal of Computer Science and Information Technology. Vol. 8, No. 4.

[45] Ren, K., Wang, C and Wang, Q. (2012). Security Challenges for the Public Cloud. IEEE Internet Computing, Vol. 16, No.1, pp. 69-73.

[46] Sah, S.K., Shakya, S. and Dhungana, H. (2014). A security management for cloud based applications and services with diameter-AAA. In Proceeding of IEEE International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). pp. 6–11.

[47] Scott, P., Paul, T. J and Susan, C.W. 2010. Identifying the Security Risks Associated with Governmental Use of Cloud Computing. Journal of Government Information, Quarterly 27, pp. 245-253.

[48] Shukla, S. (2014). Public Cloud Security Challenges and Solution. International Journal of Scientific Engineering and Research. Vol. 2, Issue 4, pp. 111-117.

[49] Siddiqui, S.; Darbari, M and Yagyasen, D. (2019). A Comprehensive Study of Challenges and Issues in Cloud Computing. In Soft Computing and Signal Processing; Springer: Singapore, 2019; pp. 325–344.

[50] Singh, J. (2017). Study on Challenges, Opportunities and Predictions in Cloud Computing. International Journal of Modern Education and Computer Science, Vol. 3, pp. 17-27.

[51] Sivasubramanian, Y.; Ahmed, S. Z and Mishra, P. V. (2017). Risk Assessment for Cloud Computing. International Research Journal of Electronics & Computer Engineering. Vol. 3, No. 2. Pp. 7-9.

[52] Srivastava, P and Khan, R. (2018). A Review Paper on Cloud Computing. International Journals of

Advanced Research in Computer Science and Software Engineering. Vol. 8, Issue 6, pp .17-20.

[53] Stoneburner, G., Goguen, A. and Feringa, A. (2002). NIST SP 800-30 Risk Management Guide for Information Technology Systems. NIST, pp. 8-26.

[54] Tabrizchi, H and Rafsanjani, M.K. (2020). A Survey on Security Challenges in Cloud Computing: Issues, Threats, and Solutions. Journal of Supercomputers. Vol. 76, pp. 9493–9532.

[55] Uri, B. (2013). Enterprise Cloud Strategy: Lessons learned from recent cloud outages. https://www.rightscale.com/blog/enterprise-cloud-strategies/lessons-learned-recent-cloud-outages. (Accessed: 23 January, 2019)

[56] Velliangiri, S.; Karthikeyan, P and Kumar, V.V. (2020). Detection of Distributed Denial of Service Attack in Cloud Computing Using the Optimization-Based Deep Networks. Journal of Experimental and Theoretical Artificial Intelligence Vol. 33, Issue3, Pp.405-424.

[57] Verma, D. K and Sharma, T. (2019). Issues and Challenges in Cloud Computing. International Journal of Advanced Research in Computer and Communication Engineering. Vol. 8, Issue 4, pp. 188-195.

[58] Wu, Y., Lyu, Y and Shi, Y. (2019). Cloud Storage Security Assessment through Equilibrium Analysis. Tinshhua Journal of Science and Technology. Vol. 26, No. 6. pp. 738-749.

[59] Xu, J., Liang, C., Jain, H. K and Gu, D. (2019). Openness and Security in Cloud Computing Services: Assessment Methods and Investment Strategies Analysis. IEEE Access, Vol. 7, pp. 29038-29050.

[60] Yang, C.; Tan, L.; Shi, N.; Xu, B.; Cao, Y and Yu, K. (2020). AuthPrivacyChain: A Blockchain-Based Access Control Framework with Privacy Protection in Cloud. IEEE Access Vol. 8, pp. 70604–70615.

[61] Yogeshwaran, S., Syed, Z. A and Ved, P. M. (2017). Risk Assessment for Cloud Computing. International Research Journal of Electronics and Computer Engineering. Vol. 3, No 2. pp. 7-9.

[62] Yunchuan, S., Junsheng, Z., Yongping, X., et al. (2014). Data Security and Privacy in Cloud Computing. International Journal of Distributed Sensor Networks. Volume 2014, http://dx.doi.org/10.1155/2014/190903. (Accessed on 17 January, 2022).

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 223-230, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1007

The Development of Electronic Module Based on Scientific Literacy on Colloidal Topic

Ramlan Silaban
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Marham Sitorus
Chemistry Study Program
Universitas Negeri Medan
Medan, Indonesia

Freddy Tua Musa Panggabean
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Elssya Manullang
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Abstract: This research is motivated by the problem of the lack of variety of teaching materials used during the teaching and learning process in schools. The reason for this research is to design electronic module teaching materials based on scientific literacy on colloidal topics. This research is a development research consisting of 4 stages, namely: define stage, design stage, develop stage, and disseminate stage. This research instrument uses validation sheets and questionnaires which are analyzed descriptively qualitatively and descriptively quantitatively. This electronic module was validated by 4 media validators, namely 2 chemistry lecturers and two chemistry teachers and material validators, namely 1 chemistry lecturer and two chemistry teachers. Based on the results of the study, the percentage of assessment on the validation of teaching materials by media experts was 89% with very feasible assessment criteria and the percentage of assessment on validation of teaching materials by material experts was 91.75% with very appropriate assessment criteria on aspects of content feasibility, language feasibility, presentation and presentation feasibility. graphic eligibility. The response of class XII IPA 7 students from SMA Negeri 7 Medan was obtained with a total percentage of 91.29%, which means students strongly agree and accept and respond to teaching materials very well.

Keywords: teaching materials; electronic module; scientific literacy; chemistry; colloid

1. INTRODUCTION

Improving the quality of human resources has a close relationship with improving the quality of education [1]. One of the ways to improve the quality of education is the use of learning media in the process. This is because the use of learning media can support and attract students' interest in participating in learning activities and the use of media in learning must be tailored to the needs of the students, such that the media utilized is appropriate for the topic delivered [2]. Technology's integration into the teaching and learning process has created a new learning environment that can help students become more interested in the subjects they are studying [3]. A means of simplifying teaching materials so that it is easier for students to understand. One of the objectives of teaching materials is to adjust the content based on the demands of the curriculum by considering the needs of students. Examples of teaching materials are books, student worksheets and learning modules.

Along with the development of technology, the shape of the module is also growing and has a positive impact. The development of the module starts from the form of a printed module to the form of an E-module. E-modules as teaching materials can contain interactive experiments and simulations combined with pictures, videos and animations. E-modules can be used in chemistry subjects because chemistry is one of the sciences that develops along with technological developments and their application in everyday life. Chemistry subjects have a goal where students have the ability to understand chemical concepts, principles, laws, and theories as well as their application and solving related problems in everyday life and technology [4].

One of the chemistry topics that requires modules as teaching materials is the topic of colloids. This is because the topic of colloids contains material that requires the help of special media to visualize the properties, formation of colloids and their application in everyday life which does not allow all to be practiced or shown directly on the grounds that it is dangerous and expensive [5].

Based on the results of an interview with one of the chemistry teachers at SMA Negeri 7 Medan, explained that the ongoing learning could not be carried out fully face to face, causing the colloid learning process to be limited and still only guided by textbooks. Limited textbooks mean that students do not have other sources of reading related to colloidal material. This is in line with the opinion that textbooks provided by the school in fact cannot be used by students on the grounds that the number of textbooks is not proportional to number of students, causing learning not take place properly [6]. The existing ones are expected to be able to improve students' scientific literacy skills in colloidal topics.

Scientific literacy is the ability to identify, understand and interpret science-related issues that a person needs to make decisions based on scientific evidence [7]. The application of the concept of literacy in the colloidal topic process is not only intended to understand a collection of facts and theories but is actually the realm of a learning process to understand and interpret phenomena and events that are relevant to daily life. That is why it is important to develop electronic science literacy-based modules on colloidal topics.

1.1 Teaching Materials

Teaching materials are all forms of materials that are systematically arranged that allow students to study independently and are designed in accordance with the applicable curriculum. Teaching materials have unique and specific properties [8]. Unique, means that the teaching materials can only be addressed to certain subjects in a particular learning process as well. Specific, means that the content of teaching materials has a purpose in the learning process [9]. Therefore, it can be concluded that teaching materials play an important role in helping teachers in teaching and learning activities that are arranged systematically in written and unwritten form in order to create an effective learning atmosphere and achieve the desired learning objectives.

1.2 Learning Modules

A module is one of the teaching materials that are packaged in a systematic and complete form which contains a set of planned learning experiences to assist students in achieving learning objectives [10]. Furthermore, the module is a collection of subject matter that is compiled in writing so that students are able to absorb the material themselves [11].

The function of the module is to overcome the weaknesses of the traditional teaching system; to increase learning motivation; to enhance the creativity of trainers in preparing individual lessons; to realize the principle of continuous progress andto realize concentrated learning [12].

1.3 E-Module

The use of teaching materials with technology is an integrated technology. E-modules are included in integrated technology because the modules are teaching materials with the help of computer technology developments. The E-module is a type of print media that can be transformed in its presentation in digital or electronic form [13]. The learning process in the E-module is designed not only centered on educators but also provides opportunities for students to construct their knowledge and skills based on independent learning [14].

1.4 Scientific Literacy

Scientific literacy is the main goal of science education. Scientific literacy is more than understanding scientific knowledge. Scientific literacy can be an access where students can ask questions, find, and make decisions that are developed from their curiosity related to their daily life experiences. Scientific literacy means that students can ask questions, find, or determine answers to questions that come from everyday experiences [15]. Scientific literacy requires knowledge of scientific concepts and theories as well as knowledge of general procedures and practices related to scientific research and scientific progress. The learning process that involves science in it can create students who have the ability to communicate, the ability to think, the ability to solve problems to the ability to master technology. Therefore, it is important to involve scientific literacy in learning because scientific literacy also has a pedagogical mission to learning activities. The pedagogical mission of scientific literacy is to produce human resources who have critical, creative, innovative and productive thinking.

The characteristics of scientific literacy have been grouped by PISA. The general classification of scientific literacy is as follows [16] :

a. Natural content and changes that occur due to human activities.

b. The process of science, the ability of students to identify scientific issues, explain natural phenomena scientifically.

c. In the context of science, science participants are able to apply science to solve real problems in daily life, technology, health and the earth and the environment.

There are four components that must be considered when developing science teaching materials in the form of science modules. The four components are science as a body of knowledge (the knowledge of science), science as a way of investigating (the investigative nature of science), science as a way of thinking and the interaction of science, technology and society [17].

a. The knowledge of science

   This category intends the text to present, discuss, or ask students to remember information, facts, concepts, principles, laws, theories, and others. Textbook materials in this category include, (a) presenting facts, concepts, principles, and laws; (b) presenting hypotheses, theories, and models and (c) asking students to recall knowledge or information.

b. The investigative nature of science

   This category denotes a text that encourages students to think and act by encouraging them to "find out." Textbook materials in this category are (a) require students to answer questions through the use of materials; (b) requires students to answer questions through the use of graphs, tables, and others; (c) require students to make calculations; and (d) involve students in thought experiments or activities.

c. Science as a way of thinking

   This category denotes that the text's purpose is to show how science in general, or a certain scientist in particular, went about "finding out." Texts in this category are such as (a) describing how a scientist experimented' (b) depicts the historical development of an idea; (c) emphasizes the empirical nature and objectivity of science, and (d) discusses evidence and implements evidence.

d. Interaction of science, technology and society

   This category is meaningful to demonstrate the consequences or implications of science on society. Texts in this category are (a) explain the benefits of science and technology to society, (b) emphasizing the negative impacts of science and technology on society, (c) explore societal concerns linked to research or technology and (d) highlight scientific and technical vocations and jobs.

1.5 Colloidal Learning

In colloidal material, the recommended approach to be used in the process is a scientific approach. This is because the 2013 curriculum uses a scientific approach so that students get an understanding to know, understand and practice scientifically related to colloid lessons. The learning model implemented in the 2013 curriculum related to colloidal material is inquiry learning. Inquiry learning is a series of learning activities in which all abilities of students are maximally involved to seek and investigate systematically, critically, and logically so that they can find their own knowledge, attitudes and skills as a form of behavior change [18]. In the process of implementing inquiry learning on colloid topics, students are involved in

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 223-230, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1007

finding the essence of colloid subject matter and the teacher only acts as a guide or facilitator in the colloid learning process.

2. METHODS

2.1 Location and Time of Research

This research was conducted at SMA Negeri 7 which is located on Jalan Timor No. 36, Gaharu, Medan City from October 2021 to January 2022.

2.2 Subject and Object of Research

The subject of this research is the development of an E-module based on Scientific Literacy on colloidal topic consisting of: 1) 3 expert validators (lecturers) in the chemistry department; 2) 2 chemistry teachers at SMA Negeri 7 Medan; 3) class XII students at SMA Negeri 7 Medan, total students are 36 students. While the object of this development research is an E-module based on Scientific Literacy on colloidal topic.

2.3 Type and Design of Research

The type of research carried out in this research is the type of research and development or Research and Development (R&D) which refers to the research design of the 4-D model development. The 4-D development model consists of 4 main stages, namely: Define, Design, Develop, Disseminate [19]. In general, the stages of research and development includes (1) the define stage, which involves identifying problemstyp and potentials in the classroom, (2) the design stage, which involves creating scientific literacy E-module, (3) the develop stage, which includes validation of the design and then revisions to the design, and (4) the disseminate stage, which is the step in which the electronic chemistry module is distributed to schools [20]. This research was carried out up to the disseminate stage. Trials on E-modules that have been designed and developed are carried out on students in class XII MIA 7.

2.4 Data Collection Techniques

The data to be collected is qualitative , namely data in the form of a description in the form of sentences. Qualitative researchers are human instruments, function in determining the focus of research, selecting sources, collecting data, assessing data quality to analyzing data and making conclusions on their research [21]. This data will contain the results of interviews with resource persons. Furthermore, data collection using a questionnaire . This data consists of answers that come from validators for products that have been developed, descriptions of the implementation of product trials and the results of student responses to products that have been developed.

2.5 Research Instruments

To obtain data in this study, non-test instruments were used. The non-test instrument was used to analyze the E-module based on scientific literacy, the validity and to determine the students' responses to the E-module based on scientific literacy. The instrument used is a checklist sheet with a Likert scale. The Likert scale is applied as one of the most basic psychometric tools that is often used in educational and social science research [22]. This study uses questionnaire that aims to collect feasibility data for the developed scientific literacy-based E-module and also a student response questionnaire to obtain response data to the E-module. This research was conducted in accordance with the procedures that have been prepared. The following in Figure 1 can be seen the research procedures carried out.

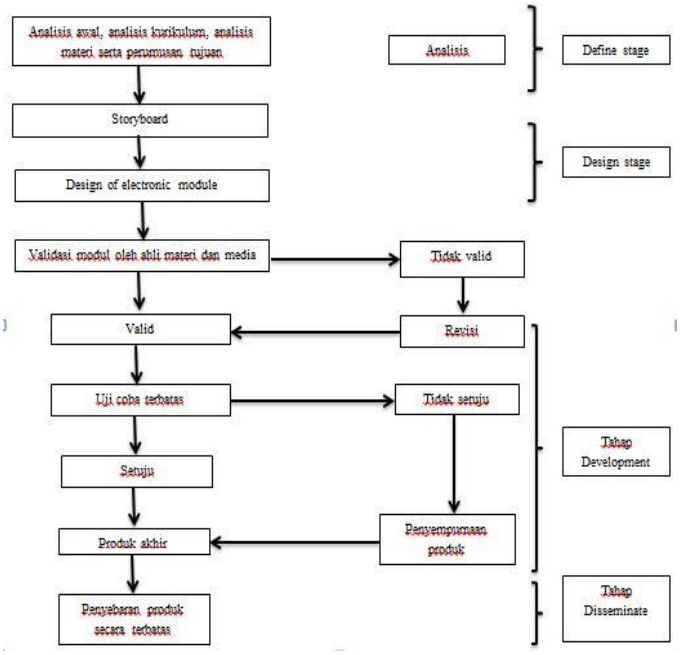


Figure. 1 Research procedure

Based on Figure 1, the procedure is described as follows.

a. Define Stage

The first stage in the research is needs analysis or is called the definition stage. The researcher analyzes the syllabus in accordance with the 2013 curriculum to obtain information about what teaching materials are suitable for the needs of students. Information about these teaching materials serves as input in the development of teaching materials, namely E-modules based on scientific literacy.

b. Design Stage

This stage is the stage of designing or compiling a draft which will be developed into an E-module based on scientific literacy. The development of these teaching materials is also based on the analysis that has been carried out which was obtained based on initial observations and interviews with teachers.

1. Selection of Teaching Materials

   The selection of teaching materials is based on the characteristics of the 2013 Curriculum High School syllabus material that is in accordance with the needs of students. This development uses E-modules as teaching materials.

2. Selection of Format

   The format selection has been adjusted to the teaching materials used. The selection of this format includes the design of the content of teaching materials or E-modules, layout design, writing fonts, images and so on.

3. Preliminary Design

This stage is designing a E-module based on scientific literacy and then consulting with the supervisor. Suggestions and inputs obtained from supervisors can be used as revision material.

c. Develop Stage

This stage produces an E-module based on scientific literacy that has been revised and then a limited trial will be conducted on the module. The development is carried out by involving relevant references and scientific literacy learning.

Product validation was carried out by media experts, material experts and chemistry teachers using the BSNP questionnaire as a validation instrument. The validator assesses based on the criteria set by BSNP which includes the feasibility of content, language, presentation and graphics. The validator also provides suggestions regarding the developed E-modules based on scientific literacy. After doing the validation, then revision of the E-module based on Science Literacy was carried out.

d. Dissemination Stage

The dissemination stage is carried out when the trial is limited and the instrument has been revised. The purpose of implementing this stage is to disseminate E-modules based on Science Literacy. This research only conducted limited dissemination. Limited dissemination is carried out by disseminating and promoting the final product of the E-module by distributing limited student response questionnaires to students at SMA Negeri 7 Medan. A limited student response questionnaires was conducted to determine the application of the E-module in learning by looking at the results of student responses then carry out product improvement.

2.6 Data Analysis

Data analysis in research intends to process data obtained from research results so that they can be accounted for and believed to be true. Data analysis is one of the requirements so that research data can be used for hypothesis testing [23]. The data analysis techniques used in this research are descriptive qualitative data analysis techniques and quantitative data analysis techniques that describe the results of the validity test and student responses. The two data analysis techniques are presented as follows.

a. Qualitative descriptive analysis

Qualitative data analysis was carried out by grouping all information in the form of input, criticism and suggestions for improvement contained in the questionnaire. The data processed in the qualitative analysis came from the results of validation and review by media experts and material experts on the E-Module based on scientific literacy on colloidal material.

b. Quantitative descriptive analysis

Quantitative descriptive analysis is done by analyzing quantitative data in the form of numbers. Quantitative descriptive analysis is used to analyze the data obtained from the questionnaire:

1. E-module Validity Analysis

The data obtained in the questionnaire will be processed by means of descriptive statistics. The rating scale used in the modified BSNP eligibility questionnaire is 1 to 4, where the lowest score is 1 and the highest score is 4. The formula used to calculate validation data based on the questionnaire is as follows:

$$\text{Percentage of validity} = \frac{Obtained\ score}{Maximal\ score} \text{ x } 100\%$$

The value the percentage of the scale of feasibility or product validity carried out by media experts, material experts and chemistry teachers is as follows.

Table 1. Criteria for the percentage of media validity [24]

Interval (%)	Qualification	Feasible Criteria
81-100	Very valid	Very worthy / not revised
61-80	Valid	Eligible/not revised
41-60	Sufficiently valid	Decent enough/needs revision
21-40	Less valid	Not worthy/revision
0-20	Invalid	Very unworthy/total revision

2. Student response questionnaire

The student questionnaire or response questionnaire was carried out after the students used the E-module. The student response questionnaires were analyzed descriptively. The category of student response assessment can be presented in the following table.

Table 2. Category of student response assessment [25]

Percentage (%)	Criteria
76-100	Strongly agree/very good
51-75	Agree/good
26-50	Disagree/bad
0-25	Strongly disagree/very bad

The formula used to calculate student response questionnaire data is as follows:

$$\text{Percentage} = \frac{Obtained\ score}{Maximal\ score} \text{ x } 100\%$$

The results of the data analysis that have been carried out will then be adjusted to agreement on numbers so that the criteria for validity and numbers can be determined. After doing this, conclusions and results from the data analysis can then be drawn.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 223-230, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1007

3. RESULT

3.1 Result of Define Stage

At the define stage or needs analysis that has been carried out includes determining and defining facts and a series of needs in the chemistry learning process at SMA Negeri 7 Medan as well as collecting initial information regarding the conditions and products to be developed. Needs analysis is done by carrying out field observations in schools.

The initial analysis was carried out by conducting interviews with chemistry teachers at SMA Negeri 7 Medan to determine the need for teaching materials to be used in the chemistry learning process at school, especially on colloidal materials used in SMA Negeri 7 the. The results of the initial analysis based on class observations and interviews with chemistry subject teachers, namely learning limitations due to the covid-19 pandemic made it difficult for teachers to provide maximum understanding to students, limited teaching materials used in the learning process and E-modules as independent teaching materials had not been used in the learning process.

Curriculum analysis that has been carried out is to find out core competencies and basic competencies related to colloidal material and to find out what materials are in chemical colloid materials that can be used as materials for making chemical teaching materials in the form of E-modules. The 2013 curriculum is used as a reference in the design phase of products, structures and components of teaching materials. At this stage, the syllabus is used as a guide.

Material analysis carried out to select relevant material from several source books and then rearrange it. Analysis of colloidal material contained in the textbooks for teachers and students from 2 different publishers. The results obtained are that the aspects contained in scientific literacy are still not fulfilled in every sub topic of colloid material discussed. The two books still do not place all aspects of scientific literacy on the 5 topics discussed. The topics are as follows: 1) colloid system, 2) types of colloids, 3) colloid properties, 4) colloid manufacture and 5) colloid application in life.

The formulation of the objectives that have been implemented serves to limit the research so as not to deviate from the initial objectives of learning. Based on the analysis that has been carried out, the purpose of the E-module is to overcome existing problems and limitations of teaching materials and increase students' interest in learning in colloid topics and develop students' scientific literacy skills.

3.2 Result of Design Stage

The teaching materials selected and developed by researchers were E-module teaching materials. Researchers used Microsoft Word application to create and design the E-module teaching materials. Then the files or teaching materials that are still in word form are converted to PDF and then published to FlipHTML5. The application can make the flip book maker more appealing, this multimedia gadget can contain files in the form of pdf, photos, videos, and animations. Background, control buttons, navigation bar, hyperlinks, and back sound are all included in flip book maker's design templates. Students can read as if they were physically reading a book since there is an animation effect that simulates physically opening a book when switching pages [26].

The steps for preparing the product design for this E-module include adjusting the core competencies and basic competencies as well as the syllabus based on the 2013 curriculum. The format used in product design is in the form of learning media for E-modules based on scientific literacy which can be compiled into a draft module. The draft of the module teaching materials was prepared according to the results of the book analysis conducted by the researcher.

3.3 Result of Development Stage

The develop stage functions to produce product that have been revised based on suggestions and input from media experts and material experts. This development stage includes a product validation test (E-module colloid based on scientific literacy) by experts and product revision. The E-module validation aims to determine the feasibility of the media and material on the E-module to be used in learning activities.

The feasibility level of the E-module teaching materials is carried out by means of validation carried out by media experts and material experts.

a. Material Expert Validation

The questionnaire used consisted of 26 assessment items with a score range of 1-4 points accompanied by 5 supporting questions. Aspects of assessment by material experts include aspects of content feasibility, aspects of language feasibility and aspects of presentation feasibility and aspects of scientific literacy. The data on the results of the assessment by material experts on this E-module can be seen in Table 3 below.

Table 3. Result material expert validation

Aspect	Validator			Average	%	Criteria Feasibility
	1	2	3			
Content feasibility	3.62	3.67	3.68	3.65	91.25	Very feasible
Language feasibility	3.33	3.83	3.5	3.55	88.75	Very feasible
Presentation feasibility	3.67	3.71	3.71	3.69	92.25	Very feasible
Scientific literacy	4	4	4	4	100	Very feasible

Based on Table 3, it can be seen that the average score of the material assessment carried out by the validator is the content feasibility aspect of 91.25%, the language feasibility aspect of 88.75%, the presentation feasibility aspect of 92.25% and the scientific literacy aspect of 100%. Based on the feasibility criteria of the validation results by the three validators, the results of the validation of the material for developing this electronic module are in the "very feasible" criteria as an E-module learning media based on scientific literacy on colloidal material.

b. Media Expert Validation

The questionnaire used consisted of 12 statement items with a score range of 1-4 points with 2 supporting questions. The assessment is based on module size, module cover design and module content design. The data on the results of the assessment by media experts on this electronic module can be seen in table 4 below.

Table 4. Result material expert validation

Aspect	Validator				Average	%	Criteria Feasibility
	1	2	3	4			
Graphic feasibility	3.89	3.86	3.39	3.08	3.56	89	Very feasible

From the table above, it can be seen that the average score of media assessment by validator 1 is 3.89 (97.25%), validator 2 is 3.86 (96.5%), validator 3 is 3.39 (84.75%) and validator 4 is 3.08 (77%).) with the average graphic aspect obtained at 3.56 with a percentage of 89%. Based on the feasibility criteria of the validation results by the four validators, the results of this development are in the "very feasible" criteria as a learning media for E-modules based on scientific literacy on colloidal material.

c. Product Revision

The product revision stage is carried out after the material validator and media validator provide suggestions. The revised e-module is then sent to the material and media validator for re-assessment. The product revision is complete if the material and media validator has stated that the following module electronics are valid. Some of the improvements implemented include the following, 1) Simplification of definitions in several parts, 2) adding an explanation of the material in the colloid sub-topic (the application of colloids in daily life), 3) improving the e-module cover, 4) changing the color or font size on the concept map and 5) changing the layout of the competency achievement indicators

3.4 Result of Dissiminate Stage

Media dissemination is a stage that is carried out when the electronic module learning media based on scientific literacy on colloidal material has been revised for the better. This stage is carried out by distributing the media in related classes along with distributing several student response questionnaires in class XII IPA 7 at SMA Negeri 7 Medan. The process of collecting data is by providing learning media, namely an electronic module based on scientific literacy on Colloidal material made using FlipHTML5, then distributing an assessment questionnaire (student response questionnaire) in the form of a link from the google form.

The questionnaire used was 22 assessment items with a score range of points 1-4. Aspects of student response include aspects of appearance, material aspects and aspects of benefits. The data on the results of the assessment by material experts on this electronic module can be seen in Table 5.

Table 5. Student Responses Results

Aspect	Percentage
Appearance	93.06 %
Material	90.12 %
Benefit	90.69 %
Total	91.29 %

Based on the results of data analysis on student response questionnaires that have been filled out by 36 students, where the number of students who chose the "strongly agree" category were 31 students, the results obtained from the criteria for the results of student responses with an average percentage of 91.29% with criteria "Strongly agree". Overall, the electronic module learning media based on scientific literacy on colloid topics does not need to be revised again. The electronic module can be used as an independent teaching material on Colloidal material that will be faced by class XI students in the second semester.

4. DISCUSSION

The research entitled the development of electronic module based on scientific literacy on colloidal topic, aims to produce electronic module teaching materials that are based on four aspects of scientific literacy. Independent teaching and get student responses to scientific literacy-based electronic modules on colloidal material. The research and development procedure used is the Research and Development (R&D) method with a 4-D model consisting of the analysis stage (define), the design stage (design), the development stage and the disseminate stage.

The first step that researchers have taken in this research is the define stage. This stage intends to find and collect problems for which solutions are sought. In this study, the needs analysis was carried out to identify and define the problems encountered in colloidal learning. In this stage, the researcher conducted initial observations by conducting interviews with one of the chemistry teachers at SMA Negeri 7 Medan, to find out the teaching materials and media used by chemistry teachers in the chemistry learning process at school, especially colloidal material. The results of the interviews that learning chemistry in the classroom is still only guided by the teacher and textbooks. Teachers still use learning media only in the form of powerpoint. The chemistry learning has also never used electronic modules as teaching materials and scientific literacy-based learning in chemistry is rarely applied so that students are less active and independent. Based on this information, the researchers got the motivation to develop an electronic module based on scientific literacy on colloidal material. This electronic module is made using FlipHTML5 which can support the module to be more interactive in its use. This electronic module was created to overcome existing problems and aims to minimize students' lack of interest in chemistry

subjects by producing scientific literacy-based teaching materials that can make it easier for students to learn both in groups and independently.

After the define stage is implemented, the next stage is the design stage. At this stage, the product, namely an electronic module based on scientific literacy on colloidal material, begins to design its components. This starts from the selection of teaching materials, namely electronic modules that aim to facilitate the learning process, as well as independent teaching materials for students because teaching materials in the form of electronic modules have never been developed at SMA Negeri 7 Medan. In fact, the module teaching material is a relevant teaching material and is a good solution when learning is still carried out on a limited basis, as it is today. At the stage of selecting teaching materials, an electronic module was developed from the 2 books that were analyzed. The next step is the preparation of the electronic module product design. What is being carried out is adjusting core competencies and basic competencies based on the 2013 syllabus and curriculum. The format used in product design is in the form of electronic module learning media that refers to the 2013 curriculum by paying attention to the four components of scientific literacy in it, namely the knowledge of science, the investigative nature of science, science as a way of thinking, and interaction of science, technology and society. At this stage, the researcher writes a draft module that is developed according to the curriculum and syllabus which is then inserted into the electronic module so that the electronic module is made in accordance with the subject matter and scientific literacy.

The next stage is the development stage. This stage includes product validation and getting student responses regarding this electronic module. This stage is carried out to determine the level of media feasibility and the feasibility of the material developed in accordance with the 2013 curriculum and the syllabus used. This was assessed using a BSNP questionnaire given to expert validators where the material validator was one lecturer in chemistry at the State University of Medan and two chemistry teachers at SMA Negeri 7 Medan. one chemistry teacher at SMA Negeri 7 Medan. Based on suggestions regarding product deficiencies provided by expert validation, it is expected to make the module better and feasible to use in the learning process.

The validation results obtained from the validator are then analyzed. The validation carried out by material experts obtained results where aspects of content feasibility, language feasibility, presentation feasibility had an average of 91.75% which met the "very feasible" criteria, also accompanied by notes and suggestions that were used as guidelines for revising the material. Furthermore, the validation carried out by media experts got the results of the feasibility of graphics having an average of 89% which met the "very feasible" criteria, also accompanied by notes and suggestions that were used as guidelines for media revision. Based on the results of validation carried out, it can be concluded that e-module based on scientific literacy on colloidal material is very feasible to use.

The results of this study are in line with research conducted by Novia F. J which showed that the Android E-Module based on scientific literacy integrated with Islamic values in the reaction rate material produced was tested valid with percentage of 91.2% (very valid) [27]. In addition, research conducted by Linda Rosita that stated that the development of problem-based learning-based electronic module teaching materials on reaction rate materials that had been tested for media feasibility was 92% and material feasibility was 91% [28]. This shows that the feasibility test of the electronic module based on problem based learning on the reaction rate material is considered very feasible.

The last thing that was carried out was distributing electronic modules along with the results of student responses to class XII IPA 7. This stage was carried out to promote the product developed so that it was accepted by users. There are three main stages in the disseminate stage, namely validation testing, packaging and diffusion and adoption. Validation testing is a product that has been revised at the development stage and then implemented on the real target, packaging, diffusion and adoption, namely product packaging is done by printing the product which is then disseminated so that it can be absorbed (diffusion) or understood by others and can be used (adopted) in the classroom related. At this stage, a limited disseminate stage is implemented, the researcher carried out a limited trial to get student responses. This is done by giving questionnaires to students. The average obtained is 91.29% with the "strongly agree" category, so that the overall scientific literacy-based electronic module on colloidal material does not need to be revised and is worthy of being used as learning media.

The results of this study of student response to electronic module in line with research that done by Rizka Annisa Rahman which that the development of an electronic module based on Problem Based Learning (PBL) on thermochemical material obtained a total student response of 92.2%, which means that students accept and respond to teaching materials very well [29].

5. CONCLUSION

Based on the results of the analysis that has been carried out in this study, it can be concluded that:

The level of validity of an electronic module based on scientific literacy on colloidal topic as teaching material for SMA/MA class XI students was declared "very feasible" by material experts getting a score of 3.67 with a percentage of validity of 91.75% and declared "very feasible" by media experts with a value of 3.56 with a percentage of validity of 89%.

The level of student response related to an electronic module based on scientific literacy on colloidal topic as teaching materials for SMA/MA students in class XII SMA Negeri 7 Medan obtained the percentage of assessments stated "strongly agree" and "very good ” with a response percentage of 91.29%.

6. REFERENCES

[1] Agustina, N. R., Rachman, F. A., & Nawawi, E. 2018. Penerapan Model Pembelajaran Kooperatif Tipe Teams Games Tournament (TGT) untuk Meningkatkan Hasil Belajar Kimia Siswa Kelas X SMA Negeri 10 Palembang. *Jurnal Penelitian Pendidikan Kimia: Kajian Hasil Penelitian Pendidikan Kimia*, 5(2): 137-146.

[2] Panjaitan, H. P., Silaban, R., Jahro, I. S., Hutabarat, W., Riris, I. D., Sudrajat, A., & Nurfajriani. 2021. Development of Innovative Chemistry Practicum Based on Multimedia Senior High School Class XI Semester II Integrated Character Education According to the 2013 Curriculum. *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, 4(2): 880-887.

[3] Purba, J., Situmorang, M., & Silaban, R. 2019. The Development and Implementation of Innovative Learning Resource with Guided Projects for the Teaching of Carboxylic Acid Topic. *Indian Journal of Pharmaceutical Education and Research*, 53(4): 603-612.

[4] Donasari, A., & Silaban, R. 2021. Pengembangan Media Pembelajaran Kimia Berbasis Android pada Materi Termokimia Kelas XI SMA. *Jurnal Inovasi Pembelajaran Kimia*, 3(1): 86-95.

[5] Sari, I. N., Saputro, S., & Ashadi. 2013. Pengembangan multimedia pembelajaran berbasis macromedia flash sebagai sumber belajar mandiri pada materi koloid kelas XI IPA SMA dan MA. *Jurnal Pendidikan Kimia (JP*K), 2(3): 152-17.

[6] Silaban, R., & Sianturi, P.A. 2021. Pengembangan Media Pembelajaran Kimia Berbasis Android pada Materi Laju Reaksi. *Jurnal Inovasi Pembelajaran Kimia*, 3(2): 191-200.

[7] Kementrian Pendidikan dan Kebudayaan. 2017. *Konsep Literasi Sains Dalam Kurikulum 2013*. Jakarta: Pusat kurikulum dan Perbukuan.

[8] Magdalena, I., Sundari, T., Nurkamilah, S., Nasrullah., & Amalia, D. A. 2020. Analisis Bahan Ajar. *Jurnal Pendidikan dan Ilmu Sosial*, 2(2): 311-326.

[9] Sihotang, R. 2014. Mengembangkan bahan ajar dalam pembelajaran ilmu pengetahuan sosial (IPS) di SD. *Jurnal Kewarganegaraan*, 23(2): 13-24.

[10] Rahdiyanta, D. 2016. Teknik Penyusunan Modul. *Academia*, 1-14.

[11] LKPP. 2015. Format Bahan Ajar, Buku Ajar, Modul, dan Panduan Praktik. Makassar: UNHAS.

[12] Hernawan, A. H., Parmasih., & Dewi, L. 2012. *Pengembangan Bahan Ajar*. Bandung: Direktorat UPI.

[13] Hutahaean, L. A., Siswandari., & Harini. 2019. Pemanfaatan E-Module Interaktif Sebagai Media Pembelajaran Di Era Digital. *Prosiding Seminar Nasional Teknologi Pendidikan* (p. 298-305). Medan: Pascasarjana UNIMED.

[14] Rini, T. A., & Cholifah, P. S. 2020. Electronic Module With Project Based Learning: Innovation of Digital Learning Product on 4.0 Era. *Edcomtech*, 5(2): 155-161.

[15] Sutrisna, N. 2021. Analisis Kemampuan Literasi Sains Peserta Didik SMA Di Kota Sungai Penuh. *Jurnal Inovasi Pendidikan*, 1(12) : 2683-2693.

[16] OECD. 2016. *PISA 2015 Result in Focus*. Paris: OECD Publishing.

[17] Chiappetta, E.L., Fillman, D.A., dan Sethna, G.H. (1991b). A Quantitative Analysis of High SchoolChemistry Textbooks for Scientific Literacy Themes and Expository Learning Aids. *Journal of research in science teaching*, 28 (10) : 939-951.

[18] Hanafiah, N., & Suhana, C. 2010. *Konsep Strategi Pembelajaran*. Bandung: PT Refika Aditama.

[19] Thiagarajan, S., Semmel, D.S. & Semmel, M. I. 1974. *Instructional Development for Training Teachers of Exceptional Children.* Minneapolis, Minnesota : University of Minnesota.

[20] Khotim, H. N., Nurhayati, S., & Hadisaputro, S. 2015. Pengembangan Modul Kimia Berbasis Masalah Pada Materi Asam Basa. *Chemistry in Education*, 4(2): 63-69.

[21] Sugiyono. 2019. Metode Penelitian dan Pengembangan: Research and Development. Bandung: Alfabeta

[22] Joshi, A., Kale, S., Chandel, S., & Pal, D. K. 2015. Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, 7(4): 396-403.

[23] Silaban, R., Panggabean, F. T. M., Hutahaean, E., Hutapea, F., & Alexander, I. 2021. Efektivitas Model *Problem Based Learning* Bermediakan Lembar Kerja Peserta Didik Terhadap Hasil Belajar Kimia dan Kemampuan Berpikir Kritis Peserta Didik SMA. *Jurnal Ilmu Pendidikan Indonesia*, 9(1): 18-26.

[24] Riduwan. 2007. Skala Pengukuran Variabel-Variabel Penelitian. Bandung: Alfabeta.

[25] Riduwan. 2009. Skala Pengukuran Variabel-Variabel Penelitian. Bandung: Alfabeta.

[26] Panggabean, F. T. M., Silitonga, P. M., & Sinaga, M. 2022. Development of CBT Integrated E-Module to Improve Student Literacy HOTS. *International Journal of Computer Applications Technology and Research*, 11(5): 160-164.

[27] Jayanti, N. F. 2020. Desain dan Uji Coba E-Modul Android Berbasis Literasi Sains Terintegrasi Nilai Islam Pada Materi Laju Reaksi. Skripsi, Kimia, Universitas Negeri Sultan Syarif Kasim Riau, Pekanbaru.

[28] Rosita, L. 2021. Pengembangan Bahan Ajar Modul Elektronik Berbasis Problem Based Learning (PBL) pada Materi Laju Reaksi. Skripsi, Kimia, Universitas Negeri Medan, Medan.

[29] Rahman, R.A. 2021. Pengembangan Modul Elektronik Berbasis Problem Based Learning Pada Materi Termokimia Menggunakan Aplikasi Kvisoft Flipbookmaker. Skripsi, Kimia, Universitas Negeri Medan, Medan.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 231-235, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1008

Data Preparation for Machine Learning Modelling

Ndung'u Rachael Njeri
Information Technology Department
Murang'a University of Technology
Murang'a, Kenya

Abstract: The world today is on revolution 4.0 which is data-driven. The majority of organizations and systems are using data to solve problems through use of digitized systems. Data lets intelligent systems and their applications learn and adapt to mined insights without been programmed. Data mining and analysis requires smart tools, techniques and methods with capability of extracting useful patterns, trends and knowledge, which can be used as business intelligence by organizations as they map their strategic plans. Predictive intelligent systems can be very useful in various fields as solutions to many existential issues. Accurate output from such predictive intelligent systems can only be ascertained by having well prepared data that suits the predictive machine learning function. Machine learning models learns from data input using the 'garbage-in-garbage-out' concept. Cleaned, pre-processed and consistent data would produce accurate output as compared to inconsistent, noisy and erroneous data.

Keywords: Data Preparation; Data pre-processing; Machine Learning; Predictive models

1. INTRODUCTION

The world is witnessing a fourth industrial revolution, which is fast-paced due to technological evolutions and advancements. Today, digital systems are been experienced in all spheres of the industries including and not limited to healthcare, education, manufacturing, entertainment, and telecommunication where there's a wealth of data. The digital systems have become sources of massive data, where insights can be extracted and analyzed for new patterns and new knowledge that may be useful in building various smart applications in the pertinent domains.

2. Data Pre-processing

Data pre-processing is an important step while developing smart systems or while extracting meaningful insights using machine learning. Data processing is sometimes used interchangeably with data preparation; however, data processing is inclusive of both data preparation and feature engineering whereas data preparation excludes feature engineering [4]. Before data preparation, there is usually need to understand the output you require from the machine model to be trained, and hence the subsequent data attributes that will shape the output. With the output in mind, the data to be collected is easily identifiable, and thus its quality and value requirements defined. This problem articulation ascertains the right steps of data preparation are followed.

The data pre-processing involves data cleaning, which involves removal of 'dirt' or noise in data, removal of missing or inconsistent data, data integration if data is sourced from multiple sources, data transformations depending on the type of raw data to what the machine learning algorithms can use as inputs, data reduction where unnecessary data is removed and only data that is required to develop an application is retained [5]. Data pre-processing makes sure that the data types to use in machine learning functions are transformed, an imposition requirement by some machine learning algorithms on data, with some having non-linear relationships that complicates how the algorithms functions [6].

2.1 DATA PREPARATION

Data preparation is the process of converting raw data through pre-processing before being used in fitting and evaluating machine learning predictive systems [6]. Machine learning models are particular to their data source, and hence the credibility of the data source and utility of the data collected is essential. It is plausible for a machine learning model to be high end model but training it with the wrong data yields the wrong information. Machine learning models operate on the "garbage in, garbage out" philosophy, and data scientists ensure the "garbage in" remains relevant, for the resultant information to be relevant. Standardizing your data entry point ensures the right information is attained at the end result. For these reasons, data collection remains an imperative part of data preparation.

Data preparation ascertains minimal errors in your data, and allows for data monitoring of any future errors. This will eventual ensure the machine learning is trained with the correct data and hence the output will be accurate. Data exploration analysis will provide a summary of your data set, and allow for necessary changes or formatting to be done. Any data source in machine learning is divide into both the training and the test data, and the technique of this division is achieved during data preparation. Additionally, data preparation helps in shaping the data to fit the requirements of the machine learning model.

Some data sets have attributes that are not well ordered for analysis. Other times, the ranges in the data sets to be compared largely vary, resulting to comparison challenges. Data transformation allows for such data sets to be transformed into good representations of the initial data source, without losing data relevancy or data integrity. Some training models accept input data in certain formats, necessitating data transformation.

In an era of big data, there is need to create better storage techniques and often times this is costly, both in terms of storing the big data, and in analyzing it. Big data analytics require complex software which is expensive. Data reduction comes in handy in compressing data into more manageable volumes while retaining its relevance and integrity. Additionally, the reduced volumes can be used in computations as a representation of the whole data set with trivial to zero

impact on the initial data source, and the output of the model. Data reduction reduces the overall cost of data analysis, and saves on the time that would have otherwise been employed in future data processing.

The main four steps for data preparation are data collection, data cleaning, data transformation and data reduction.

2.2 DATA COLLECTION

Data collection is the initial stage of data preparation, and it involves deciding on the data set depending on the expected output of the machine model to be trained. Essentially, collection of the right data set ascertains the right data output. Data collection consists of data acquisition, data labeling, data augmentation, data integration and data aggregation.

2.2.1 Data acquisition.

Data acquisition involves identifying the data source, defining the methodology of collecting the data, and converting the collected data into digital form for computation. The data source can be primary, where data is obtained straight from the persons, objects or processes being studied. When your data In this stage, exploratory data analysis (EDA) is used, and it is a technique that aims at understanding the characteristics and attributes of the data sets [12]. It aids in the data scientist becoming more familiarized with the data collected. In exploratory data analysis, statistical tools and techniques are applied in building hypothesis source is a party that had previously collected data, it is termed as a secondary source. Methodology of data collection varies depending on the expected output. Statistical tools and techniques are applied in both the collection of qualitative and quantitative data.

2.2.2 Data labelling

As machine learning advances, there is development of deep learning techniques which have automated the generation of features from data sets, and hence the requirement of high volumes labelled data [7]. Data labelling is the process through which the data models are trained through tagging of data samples. For instance, if a model is expected to tell the difference between images of cats and dogs, it will be initially introduced to images of cats and dogs, which are tagged as either cats or dogs. This is done manually, though often with the aid of a software. This part of supervised learning allows the model to form a basis of future learning. The initial formation of a pattern in both the input and output data, defines the requirements of the data to be collected. Therefore, before data collection is initialized, there is need to delineate the data parameters and the intended information to be retrieved from the data.

2.2.3 Data augmentation

Data augmentation is a data preparation strategy that is used in increasing data diversity for deep learning model training [8]. It involves construction of iterative optimization with the aim of developing new training data from already existing data. It allows for the introduction of unobserved data or introduction of variables that are inferred through mathematical models [9]. While not always necessary, it is essential when the data being trained is complex and the available volume of sampled data is small. Data augmentation saves the problem of limited data and model overfitting [10].

2.2.4 Data aggregation

Data aggregation is a technique of reducing the volume of data though grouping. This grouping is usually of a single attribute. For instance, when one has a data set with the attribute time organized in days over a given time series, one can aggregate the data into monthly groups which eases dealing with the time attribute. It aids in reducing the broadness of a given attribute without tangible losses during future data manipulation [10].

2.3 DATA CLEANING

Data cleaning, also referred to as data cleansing is the technique of detecting and correcting errors and inaccuracies in the collected data [11]. Data is supposed to be consistent with the input requirement of the machine learning model. The main activities in data cleansing involve the fine-tuning of the noisy data and dealing with missing data. It aids in ensuring the collected data set is comprehensive and any errors and biases that may have arose in data collection have been eliminated. This includes the detection of outliers within the data set; both for the numerical and the non-numerical data sets.

2.3.1 Exploratory Data Analysis

on the information that can be attained from the collected data, and sometimes involves data visualization. Data visualization allows for the understanding of data properties as skewness and outliers.

Exploratory data analysis is mainly done on the statistical manipulation software. The graphical techniques allow for understanding the distribution of the data set, and the statistical summary of all attributes. EDA allows for future decisions such as the data cleansing techniques to be used, what data transformations are necessary and whether data reduction is necessary and if yes, what is technique to use. Exploratory data analysis is a continuous process all through data preparation.

2.3.2 Missing Data

While it is important to ascertain during data collection that all the attributes of the data sets have their real value collected, data sometimes has some of the attributes with missing values, which makes it hard to use as input in machine learning models. As so, different techniques have been outlined on how to deal with missing data. Data manipulation platforms as python and R statistics have some of these techniques of dealing with missing data embedded in them. The best technique usually varies with the data set, and hence after data assessment in the exploratory data analysis, one can easily select the best technique for missing data imputation.

2.3.2.1 Deductive Imputation

Deductive imputation follows the basic rule of logic, and is hence the easiest imputation, however, the most time consuming. Even so, its results are usually highly accurate. For instance, if student data indicates that the total number of students is 10, and the total number of examinations papers is 10, but there is a paper with a missing name and John has no marks recorded, logic dictates the nameless paper is John's. However, deductive imputation is not applicable in all types of data sets [13].

2.3.2.2 Mean/Median/Mode Imputation

This imputation uses statistical techniques where the central measures of tendency within a certain attribute are computed and the missing values replaced with the computed measure of central tendency, may it be mean, mode or the median of that attribute [13]. This technique is applied in numerical data sets,

and the impact on the output or later computations is trivial. Data manipulation platforms as python and R statistics have techniques of dealing with missing data embedded in them.

2.3.3 Noisy Data.

Presence of noisy data can have substantial effect on the output of a machine model. It negatively impacts on prediction of information, ranking results, and the accuracy in clustering and classification [14]. Noisy data includes unnecessary information in the data, redundant data values and duplicates or pointless data values. These result from faultiness in collection of data, problems that may result from data entry, problems that occur from data transfer techniques applied, uneven naming conventions of the data and sometimes it may arise from technology restrictions, as in the case of unstructured data. Noisy data is eliminated through.

2.3.3.1 Binning Method

This involves arranging data into groups of given intervals, and is used in smoothening ordered data. The binning method relies on the measures of central tendency and it is done in one of three ways. Smoothing by bin means, smoothing by bin median and smoothing by bin boundary.

2.3.3.2 Regression

Linear Regression is a statistical and supervised machine learning technique, that predicts particular data based on existing data [15]. Simple linear regression is used to compute the best line of fit based on existing data, and hence outliers in the data can be identified. To attain the best line fit, there is development of the regression function based on the prior collected data. However, it is important to note that though in some data sets, extreme outliers are considered noisy data, the outliers can be essential to the model.

For instance, if an online retailer company has its market within countries in Europe and trivial market in the United States, the United States may be considered an extreme outlier, and hence noisy data. However, a machine learning model may realize that though a very small number of the Americans use the online platform, they bring in more revenue than some of the countries in Europe. Simple linear regression uses one independent variable whereas multiple linear regression uses more than one independent variable in its computations.

2.3.3.3 Clustering

Clustering is in the unsupervised machine learning category and it operates by basically grouping the collected data set into clusters, based on their attributes (Gupta & Merchant, 2016). In clustering, the outliers in the data may fall within the clusters, and in the case that they are extreme outliers they fall outside the clusters. To understand the effect of clustering, data visualization techniques are used "Clustering methods don't use output information for training, but instead let the algorithm define the output" [17]. There are different techniques used in clustering.

In K-means clustering, K is the number of clusters to be made, and to do this the algorithm randomly selects K number of data points from the data set. These K data points are called the centroids of the data, and every other data point in the data set is assigned to the closest centroid. This process is repeated for all the new K data sets created, and the process iterated until the centroids become constant, or fairly constant. This is called the point at which convergence occurs. The Density-Based Clustering of Applications with Noise (DBSCAN) is used in data set smoothing.

2.4 DATA TRANSFORMATION

Data transformation involves shifting the cleansed data from one format to the next, from one structure to the next, or changing the values in the cleansed data set to meet the requirements of the machine learning model [18]. The simplicity of the data transformation is highly dependent on the required data for input, and the available data set. Data transformation involves:

2.4.1 Normalization

Normalization is a technique for data transformation that is applied in numeric values of columns when there is for a common scale. This transformation is achieved without loss of information, but only changing how it is represented. For instance, in a data set with two columns that have different scales such as one with values ranging from 100 to 1,000 and another column with a value range of 10,0000 to 1,000,000 there may arise a difficulty in the event that the two columns have to be used together in machine learning modelling. Normalization finds a solution by finding a way of representing the same information without loss of distribution or ratios from the initial data set [19].

It is imperative to note that while normalization is only necessitated by the nature of some data sets, other times it is demanded by the machine learning algorithms being used. Normalization uses different mathematical techniques such as z-score in data standardization. The technique picked is usually decided depending on the nature and characteristics of the dataset. Therefore, it is decided at the exploratory data analysis stage.

2.4.2 Attribute selection

In this transformation, latent attributes are created based on the available attributes in the data set to facilitate the data mining process [18]. The latent attributes created usually have no impact on the initial data source, and therefore can be ignored afterwards. Attribute transformation usually facilitates classification, clustering and regression algorithms. Basic attribute transformation involves decomposition of the available attributes through arithmetic or logical operations. For instance, a data set with a time attribute given in months, can have its month attribute decomposed to weeks, or aggregated to years depending on the requirements.

2.4.3 *Discretization*

In data transformation by discretization, there is creation of intervals or labels, and eventual mapping of the all data points to the created data intervals or labels. The data in question is customarily numeric data. There are different statistical techniques used in discretization of data sets. The binning method is used on ordered data, where the data is creation of data intervals called bins where all the data points are mapped into. In data discretization by histogram analysis, histograms are used in dividing the values of the attribute into disjoint ranges where all other data points are mapped to. Both binning and histogram analysis are unsupervised data discretization methods.

In data discretization by decision tree analysis, the algorithm picks the attribute with the minimum entropy, and uses its minimum value as the point from which it, in iterations, partitions the resulting intervals till it attains as many different groups as possible [20]. This discretization is hierarchical hence its name. To use an analogy, it's like dividing a room into two equal parts, and continuously dividing the resulting partitions into two other equal parts. Only in this case, the room has multi-varied contents and we want each different content in

its own space at the end of the partitioning. This discretization technique uses a top-down approach and is a supervised algorithm.

Data discretization by correlation analysis is highly dependent on mathematical tools and it applies a bottom-up approach, unlike decision trees [20]. It maps data points to data intervals by the best neighboring interval for each data point, and merging the intervals. It then recursively repeats the process to create one large interval. It is a supervised machine learning methodology.

2.4.4 Concept Hierarchy Generation

In concept hierarchy data transformation, there is mapping of low-level concepts within the attributes to higher level concepts [21]. Most of these concepts are normally implied in the initial data set, and hence the technique is embedded in statistical software. It follows a bottom up approach. For instance, in the location dimension, cities can be mapped to their states, their provinces, their countries and eventually their continents.

.

2.5 DATA REDUCTION

With the advancement of trends in information technology and the exponential growth of internet of things, there has been an eventual precipitous increase in the volumes of available data. This is a huge benefit to machine learning as the availability of big data for training the models ascertains accuracies in the outputted information from such models. Nonetheless, handling and analyzing these enormous volumes of data is a big challenge, hence the need for data reduction techniques. Data reduction reduces the cost of analyzing and storing these volumes of data by increasing storage efficiency. The different techniques used in data reduction include.

2.5.1 Data cube aggregation

A data cube is an n-dimensional array that uses mathematical tensors to represent information. the online analytical processing (OLAP) cube stores data in a multidimensional form, which occupies lesser storage space compared to a unidimensional storage technique [22]. To access data from the OLAP cube, the Multidimensional expressional (MDX) query language is used. The query language includes the roll-up, drill-down, slice and dice and pivot operations. These operations allow access to the required attributes of the data from the cube, without removing the data from the data cube, hence saving on space.

2.5.2 Attribute subset selection

Attribute subset selection, also known as feature selection is a part of feature engineering and it involves the discovery of the smallest possible subset of attributes that would yield the same results or closest to the same results on data mining, as when using all the attributes [23]. This technique ensures that only what is completely necessary from the initial data set is used in the modeling. This simplifies detection of insights, patterns and information from the data set while saving on analysis and storage costs.

2.5.3 Numerosity reduction

In numerosity reduction data reduced and made feasible for analysis through replacement of the original data with a model of the data that preserves the integrity of the initial data [24]. Two statistical method are used in the creation of the representational model. In the parametric method, regression and log-linear methods are sued in the development of the representational model. Non-parametric methods encompass the use of clustering, sampling, use of histograms and data cube aggregation to represent the whole data population, during computations and storage.

3. POSSIBLE BIASES IN DATA PREPERATION

Bias in the data to be trained in the machine learning model leads to consequent wrong information output. It is imperative to identify the source of any bias in your data set during data preparation and eliminate the bias [25]. Sample bias occurs at data collection where the selected data sample is not the right representation of the population under study, hence it is also called selection bias. For instance, an iris scan recognition trained entirely on the iris scans of Africans will not efficiently identify eyes of the white population.

Exclusion bias is common in the data cleansing stage where there is deletion, or misrepresentation of a part of the data, leading to it being excluded in the model training. Measurement bias occurs either during data collection, where the system of collecting input data is not the same as that of collecting output data. Additionally, it occurs during data labelling, where non-uniform data labelling results to faulty predictions from the machine learning model. Recall bias also occurs at the data labelling stage, where the labelling is non-consistent [25].

Observer bias is data fallacy where the person dealing with the data assumes the observation to be wat they expected, as opposed to the real observation. Data scientists and researchers are encouraged to operate on an objective rather than subjective approach to avoid this bias [19]. Another is racial bias, and the best example of this bias in talk balk engines, where the model was largely trained on the voice data of the white population, and hence it hardly recognizes the voice of the black data population [19]. Association bias occurs when a data set has created an implicit association between attributes. The main association bias is the gender bias, as in the case where a system is trained with all school principals being males, and hence eventually disqualifies the plausibility of a female school principle [25].

4. CONCLUSION

Many machine learning predictive systems and models are affected by the kind of data that is used as input of the models. Results of the predictive models are determined by the machine learning algorithm function and the kind of data input. Biased data will produce biased results. Equally, ‘dirty’ data will produce wrong results or output that cannot be relied upon.

It’s imperative to have clean data to fit in the machine learning models so as to have the models learn correctly and predict accurately. There is high chance that inaccurate results from machine learning models are caused by improperly prepared input data. Therefore, for ensuring the explainability and reliability of machine learning predictive models that are used to develop intelligent systems, clean prepared data is significant.

Digital data sources such as internet of things which is a major source of real-world data have noisy, inconsistent and missing data, which when used in predictive modelling using machine learning functions can result to erroneous and inaccurate results. Removal of such inconsistencies in input data cannot be overemphasized. Clean data which is formatted and organized to the required standard of the machine learning function goes a long way in contributing towards better machine learning models with reliable results. There is more to data preparation than has been included on this work. In future, we look to define different types of data and their various pre-processing methods.

5. ACKNOWLEDGMENTS

My thanks to all authors whom I have referenced here below for their research works, which was insightful and helped to compile above findings.

6. REFERENCES

[1] applications and research directions." *SN Computer Science* 2, no. 3 (2021): 1-21.

[2] Altexsoft. (2018, June 16). Preparing Your Dataset for Machine Learning: 8 Basic Techniques That Make Your Data Better. Retrieved on July 29, 2020 from: *https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/*

[3] Bengfort, B., & Kim, J. (2016). Data analytics with Hadoop: an introduction for data scientists. " O'Reilly Media, Inc.".

[4] El-Amir, H., & Hamdy, M. (2020). Data Wrangling and Preprocessing. In Deep Learning Pipeline (pp. 147-206). Apress, Berkeley, CA. Retrieved from: *https://doi.org/10.1007/978-1-4842-5349-6_*

[5] García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing.

[6] Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

[7] Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*. Retrieved from: *https://ieeexplore.ieee.org/abstract/document/8862913*

[8] Ho, D., Liang, E., Liaw. R., (2019, June 7). 1000x Faster Data Augmentation. Berkeley Artificial Intelligence Research. Retrieved on July 29, 2020.

[9] Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340.

[10] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60. Retrieved from: https://doi.org/10.1186/s40537-019-0197-0

[11] Murata, K., Noda, H., & Haraguchi, M. (2017). U.S. Patent No. 9,558,151. Washington, DC: U.S. Patent and Trademark Office. Retrieved from: *https://patents.google.com/patent/US9558151B2/en*

[12] Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265-276. Retrieved from: *https://doi.org/10.1016/j.hrmr.2016.08.003*

[13] Van der Loo, M., & de Jonge, E. (2017). deductive: Data Correction and Imputation Using Deductive Methods. R package version 0.1, 2.

[14] Gupta, S., & Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science*, *161*, 466-474. Retrieved from: *https://doi.org/10.1016/j.procs.2019.11.146*

[15] Elgabry, O. (2019, March 1st). The Ultimate Guide to Data Cleaning. Retrieved on July 27, 2020 from: *https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4*

[16] Gupta, A., & Merchant, P. S. (2016). Automated lane detection by k-means clustering: a machine learning approach. *Electronic Imaging*, *2016*(14), 1-6. Retrieved from: *https://doi.org/10.2352/ISSN.2470-1173.2016.14.IPMVA-386*

[17] Castañón, J. (2019, May 2nd). 10 Machine Learning Methods that Every Data Scientist Should Know. Retrieved on 26th July 2020, from*: https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9*

[18] Malik, K. R., Ahmad, T., Farhan, M., Aslam, M., Jabbar, S., Khalid, S., & Kim, M. (2016). Big-data: transformation from heterogeneous data to semantically-enriched simplified data. *Multimedia Tools and Applications*, *75*(20), 12727-12747. Retrieved from: *https://doi.org/10.1007/s11042-015-2918-5*

[19] Microsoft. (2020, April, 7th). Bias in Machine Learning. Retrieved on July 31, 2020 from: *https://devblogs.microsoft.com/premier-developer/bias-in-machine-learning/*

[20] Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., ... & Herrera, F. (2016). Data discretization: taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *6*(1), 5-21. Retrieved from: *https://doi.org/10.1002/widm.1173*

[21] Swamy, M. K., & Reddy, P. K. (2020). A model of concept hierarchy-based diverse patterns with applications to recommender system. *International Journal of Data Science and Analytics*, 1-15. Retrieved from: *https://doi.org/10.1007/s41060-019-00203-2*

[22] Shen, H., Zhang, M., & Shen, J. (2017). Efficient privacy preserving cube-data aggregation scheme for smart grids. IEEE Transactions on Information Forensics and Security, 12(6), 1369-1381. Retrieved from: https://ieeexplore.ieee.org/document/7828093

[23] Demisse, G. B., Tadesse, T., & Bayissa, Y. (2017). Data Mining Attribute Selection Approach for Drought Modeling: A Case Study for Greater Horn of Africa. *arXiv preprint arXiv:1708.05072*. retrieved from: *https://arxiv.org/ct?url=https%3A%2F%2Fdx.doi.org%2F10.5121%2Fijdkp.2017.7401&v=2a6e454a*

[24] Deepak, J. (n.d.). Numerosity Reduction in Data Mining. Retrieved on July 25, 2020 from: *https://www.geeksforgeeks.org/numerosity-reduction-in-data mining.*

[25] Liam, H. (2020, July 20th). 7 Types of Data Bias in Machine Learning. Retrieved on July 31, 2020 from: *https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/*

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 236-240, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1009

Improving Students' Chemical Literacy Ability on Equilibrium Material Using Chemical Literacy-Based Teaching Materials

Ani Sutiani
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Jamalum Purba
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Freddy Tua Musa Panggabean
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Asep Wahyu Nugraha
Chemistry Study Program
Universitas Negeri Medan
Medan, Indonesia

Ricky Andi Syahputra
Chemistry Study Program
Universitas Negeri Medan
Medan, Indonesia

Abstract: Trends in science education policy in the 21st century emphasize the importance of scientific literacy as a transferable outcome. This study aims to analyze the use of chemical literacy-based teaching materials in equilibrium material to improve students' chemical literacy ability. The results showed that the chemical literacy ability of students who were given chemical literacy-based teaching materials was higher than students who were given general chemistry teaching materials. The biggest difference in students' chemical literacy ability is at the nominal level of scientific literacy, which is 8.24 points; followed at the conceptual scientific literacy level of 6.74 points and at the functional scientific literacy level with a difference of 5.61 points.

Keywords: teaching materials, chemical literacy, equilibrium

1. INTRODUCTION

Science education has an important role in preparing quality human resources in the face of globalization. The process and learning of science can produce quality human beings by showing scientific awareness (scientific literacy) and high-level thinking skills that can create human resources capable of critical thinking, creative thinking, decision making and problem solving [1].

Chemistry is included in the science family and is one of the branches of natural science that includes concepts, rules, laws, principles, and theories [2]. The Chemistry Education Study Program at the State University of Medan has a strong commitment to aligning the chemistry learning process with technological advances according to stakeholder needs.

Chemistry is built on a product, process, and scientific attitude. Chemistry cannot be learned simply through reading, writing, or listening. Mastery of chemistry is measured through the ability to master a collection of chemical knowledge and skills to do scientific work [3].

The development of digital technology affects various aspects of education including learning strategies. The need for more flexible access to time, speed, methods and efficiency in learning creates innovative learning strategies that involve ICT. ICT-supported learning allows students to learn anything, anytime and anywhere is an advantage that facilitates the learning process [4].

The big challenge for the ideal education process is not only to prepare the nation's generation that is able to live today, but the generation that is equipped with the ability to live in the future. Challenges in the global era are increasingly complex and require problem solving with a critical mindset and full of creativity [5]. However, the main problem in learning in formal education (school) today is the low absorption capacity of students. In a more substantial sense, that the learning process until today still gives teacher centered and does not provide access for students to develop independently through discovery in their thinking process [6].

The facts show that in the learning process, especially in science learning, students tend to memorize concepts, theories, and principles without interpreting the acquisition process. As a result, students become less trained to think and use their reasoning power in understanding natural phenomena that occur or when facing problems [7].

Trends in science education policy in the 21st century emphasize the importance of scientific literacy as a transferable outcome [8]. In the 21st century, literacy ability are not only limited to the ability to read, listen, write and speak orally, but more than that, literacy ability are emphasized on literacy ability that are connected to one another in the current digital era [9].

Scientific literacy is defined as the ability to use scientific knowledge, identify questions and draw fact-based conclusions to understand the universe and make decisions about changes that occur due to human activities [1]. Scientific literacy is one of the parameters in determining the human development index which is strongly influenced by the quality of education [10].

Chemical literacy is related to how students can appreciate nature by utilizing the science/chemistry and technology they master. People who have chemical literacy understand the basic concepts of chemistry, can explain phenomena and solve problems in life using their understanding of chemistry, understand chemical innovations in social life and have an interest in chemistry [11]. Chemical literacy can be used as a forum for students to train high-level thinking where students relate to everyday phenomena [12].

Indonesia's scientific literacy ability is in the low category based on the 2015 *Programme for International Student Assessment* (PISA) study report. For scientific literacy, Indonesian students are ranked 62 out of 70 countries with a score of 403; for mathematical literacy, Indonesian students are ranked 63rd out of 70 countries with a score of 386 even for reading literacy is ranked 64th out of 70 countries with a score of 397 [1]. These results show that the average scientific literacy ability in Indonesia is only able to recognize basic facts, but has not been able to communicate and relate these abilities to various scientific topics, especially to apply concepts in life.

Many factors that can affect the low ability of chemical literacy include the education system, models, approaches, methods, learning strategies used, learning resources, learning styles and learning infrastructure. Therefore, the development of chemical literacy ability is not only influenced by the learning model but also the teaching materials used. This is in accordance with the research of Rusilowati and Safitri et al which showed that the use of teaching materials containing scientific literacy can improve scientific literacy ability [13] [14]. However, in reality the teaching materials used have not been supported to develop scientific literacy ability. This is in accordance with the results of the study [15] which shows that the science textbooks used do not contain a balanced component of scientific literacy. In addition, the presentation of the material used emphasizes the presentation of facts, concepts, principles, laws, theories and models and places more emphasis on the stage of remembering information through the questions presented [16].

Equilibrium material is one of the studies in a compulsory subject in the Chemistry Education Study Program, Chemistry Department, FMIPA, State University of Medan (UNIMED). Equilibrium material is studied in general in the Chemical Basics course and further studied in the special Kinetics and Equilibrium course. Equilibrium material discusses the concept of equilibrium which includes chemical potential, reaction isotherms, free energy and its relation to equilibrium constants and phase equilibrium which includes the Clausius clayperon equation, phase diagrams of uner systems, binary systems and their relation to azeotropic concepts and distillation principles, and ternary system diagrams, also its application in chemical systems that can be applied in life.

To improve students' abilities in learning chemistry, it is always necessary to change or innovate continuously so that they can achieve predetermined learning objectives, one of which is the ability of graduates who are sensitive to very fast information, so that ability are needed to be selective in choosing information according to what is needed. needed. Literacy ability can be trained through the educational process using a literacy-based learning model.

Scientific literacy develops through well-structured learning activities using targeted construction teaching materials in accordance with the objectives. Teaching materials that can be used by students as a source of independent learning have an important role in improving and developing abilities, including student literacy ability [5]. The existence of teaching materials to improve chemical literacy ability is expected to provide an optimum influence in the learning process to train students to find scientific knowledge independently [17], identify questions and analyze the meaning of the acquired knowledge [18].

2. METHOD

This research is a quasi-experimental research using Non-Equivalent Control Group Design which aims to determine the effect of applying learning using chemical literacy-based teaching materials to increase students' chemical literacy ability on equilibrium material. In this design, there are two groups of UNIMED Chemistry Education study program students who are used as research subjects, namely one group who gets learning using general teaching materials (control) and the other group is given learning using chemical literacy-based teaching materials (experiments).

The instrument used in data collection was a test of chemical literacy ability on equilibrium material that had met the test quality criteria including validity, reliability, level of difficulty and discriminating power of test items. Indicators of chemical literacy ability in this study are based on the level of chemical literacy, namely nominal scientific literacy, functional scientific literacy, and conceptual scientific literacy.

The research procedure was carried out through several stages, including: 1) the initial stage, namely giving a chemical literacy ability test (pretest) to determine students' initial abilities; 2) the second stage is the learning process where control class students are given learning using general teaching materials used while experimental class students are given learning using chemical literacy-based general teaching materials; 3) the final stage is giving a chemical literacy ability test (posttest) to determine students' chemical literacy ability after being given learning.

The data analysis technique used is descriptive analysis technique and inferential technique. Descriptive analysis technique was used to describe the data including the lowest, highest, average (mean) and standard deviation values. Inferential statistical techniques were used to test the research hypotheses using the t-test of two sample groups. Before testing the hypothesis, prerequisite tests were first carried out on the data using the normality test and homogeneity test.

3. RESEARCH RESULT

3.1 Data Description

Data on students' chemical literacy ability on equilibrium material were obtained based on the results of the pretest and posttest

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 236-240, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1009

Table 1. Data of initial chemical literacy ability (pretest)

Class	Level	Min	Max	Mean
Control	Nominal scientific literacy	15	57	45.36
	Functional scientific literacy	12	56	33.44
	Conceptual scientific literacy	6	50	23.13
Experiment	Nominal scientific literacy	18	58	48.13
	Functional scientific literacy	6	55	31.41
	Conceptual scientific literacy	4	50	22.81

Table 1, shows that the average students' initial ability (pretest) in the control class and the experimental class at the nominal level of scientific literacy are 45.36 and 48.13, respectively; at the level of functional scientific literacy are 33.44 and 31.41, respectively; and the level of conceptual scientific literacy is 23.13 and 22.81, respectively. These results indicate that students' initial literacy ability on equilibrium material based on the results of the pretest are still very low at the nominal scientific literacy level, functional scientific literacy, and conceptual scientific literacy level.

Table 2. Data of chemical literacy ability (posttest)

Class	Level	Min	Max	Mean
Control	Nominal scientific literacy	50	90	81.88
	Functional scientific literacy	40	80	74.69
	Conceptual scientific literacy	40	68	63.38
Experiment	Nominal scientific literacy	55	92	90.12
	Functional scientific literacy	46	90	80.30
	Conceptual scientific literacy	41	84	70.13

Table 2, shows the average chemical literacy ability of students (posttest) at the nominal scientific literacy level, for the control class it is 81.88 (competent) while for the experimental class it is 90.12 (very competent). At the level of functional scientific literacy, the control class obtained an average of 74.69 (competent enough) while for the experimental class an average of 80.30 (competent) was obtained. At the conceptual scientific literacy level, the control class obtained an average of 63.38 (less competent) while for the experimental class, the average obtained was 70.13 (competent enough).

3.2 Result of Normality Data Test

The normality test of students' chemical literacy ability data (pretests and posttests) was analyzed by Chi-square test (χ^2) at the significance level $\alpha = 0.05$.

Table 3. Result of normality test of pretest data

Class	Level	χ^2_{count}	χ^2_{table}
Control	Nominal scientific literacy	9.16	11.07
	Functional scientific literacy	8.68	
	Conceptual scientific literacy	5.95	
Experiment	Nominal scientific literacy	7.93	
	Functional scientific literacy	8.55	
	Conceptual scientific literacy	6.32	

Table 3, shows that the results of the normality test of the students' initial literacy ability data (pretest) for each class obtained the value of $\chi^2_{count} < \chi^2_{table}$ so that it can be concluded that the pretest data from each group has a normally distributed data distribution.

Table 4. Result of normality test of posttest data

Class	Level	χ^2_{count}	χ^2_{table}
Control	Nominal scientific literacy	10.16	11.07
	Functional scientific literacy	9.23	
	Conceptual scientific literacy	9.79	
Experiment	Nominal scientific literacy	5.95	
	Functional scientific literacy	9.70	
	Conceptual scientific literacy	10.45	

Table 4, shows that the results of the normality test of students' chemical literacy ability data (posttest) for each class obtained a value of $\chi^2_{count} < \chi^2_{table}$ so it can be concluded that the chemical literacy ability data (pretest) from each group has a data distribution that is normally distributed.

3.3 Result of Homogeneity Data Test

The homogeneity test of the data is intended to determine the difference in data variance from each group. The homogeneity of the data was analyzed using Fisher test (F-Test) at the significance level $\alpha = 0.05$.

Table 5. Result of homogeneity test of pretest data

Level	Class	(S^2)	F_{count}	F_{table}
Nominal scientific literacy	Control	170.56	1.51	1.82
	Experiment	169.25		
Functional scientific literacy	Control	152.45	1.26	
	Experiment	135.74		
Conceptual scientific literacy	Control	144.25	1.01	
	Experiment	95.54		

Table 5, shows that the results of the homogeneity test of the initial literacy ability data (pretest) at each level obtained the value of $F_{count} < F_{table}$ so it can be concluded that the initial literacy ability data (pretest) is at the nominal scientific

literacy level, functional scientific literacy, and conceptual scientific level. literacy has the same variance (homogeneous).

Table 6. Result of homogeneity test of posttest data

Level	Class	(S^2)	F_{count}	F_{table}
Nominal scientific literacy	Control	15.36	1.75	1.82
	Experiment	10.48		
Functional scientific literacy	Control	91.36	1.73	
	Experiment	52.81		
Conceptual scientific literacy	Control	182.34	1.28	
	Experiment	141.94		

Table 6, shows that the results of the homogeneity test of students' chemical literacy ability data (posttest) at each level obtained the value of $F_{count} < F_{table}$ so that it can be concluded that the chemical literacy ability data (posttest) of students on balance material is good at the nominal level of scientific literacy, functional scientific literacy, as well as the level of conceptual scientific literacy have the same variance (homogeneous).

3.4 Result of Hypothesis Test

After the data analysis requirements were met both normality and homogeneity of the data, then hypothesis testing was carried out using the t-test of two sample groups at the significance level $\alpha = 0.05$.

Table 7. Result of hypothesis test

Level	t_{count}	t_{table}	Criteria
Nominal scientific literacy	5.598	2.013	Significant
Functional scientific literacy	3.381		Significant
Conceptual scientific literacy	4.985		Significant

Table 7, shows that the results of hypothesis testing on students' chemical literacy ability data (posttest) for each level obtained the value of $t_{count} > t_{table}$ so that statistically the hypothesis is accepted and it is concluded that students' chemical literacy ability are given learning using teaching materials based on chemical literacy materials the balance is higher than the chemical literacy ability of students who are given learning using general chemistry teaching materials both at the nominal scientific literacy level, functional scientific literacy, and conceptual scientific literacy level.

Table 8. Differences in chemical literacy ability

Level	Mean (Class)		
	Control	Experiment	Differences
Nominal scientific literacy	81.88	90.12	8.24
Functional scientific literacy	74.69	80.30	5.61
Conceptual scientific literacy	63.88	70.13	6.74

Table 8, shows the average difference in students' chemical literacy ability between the control class and the experimental class. The biggest difference in students' chemical literacy ability is at the nominal level of scientific literacy, which is 8.24 points; followed at the conceptual scientific literacy level of 6.74 points and at the functional scientific literacy level with a difference of 5.61 points. The difference in chemical literacy abilities of students in the control class and the experimental class is in line with the distribution of the components of scientific literacy that is good and balanced in the chemical literacy-based teaching materials used. The component of science as the body of knowledge is the component that should be found in most textbooks because it contains: (1) facts, concepts, principles and laws; (2) hypotheses, theories and models; and (3) asking students to remember knowledge and information.

4. CONCLUSION

The results showed that there were differences in chemical literacy ability between experimental class students and control class students. The results of hypothesis testing show that the chemical literacy ability of students who are given learning using teaching materials based on chemical literacy in equilibrium material (experiments) is higher than students who are given learning using general chemistry teaching materials (control). The biggest difference in students' chemical literacy ability is at the nominal level of scientific literacy, which is 8.24 points; followed at the conceptual scientific literacy level of 6.74 points and at the functional scientific literacy level with a difference of 5.61 points.

5. ACKNOWLEDGEMENTS

We would like to thank LPPM Universitas Negeri Medan for funding our research and all participans and supervisors that contributed to the work in this study.

6. REFFERENCES

[1] A. Sutiani, Zainuddin, A. Darmana, and F. T. M. Panggabean, 2020. The Development of Teaching Material Based on Science Literacy In Thermochemical Topic. *Journal of Physics: Conference Series 1462*, pp. 1–6, doi: 10.1088/1742-6596/1462/1/012051.

[2] F. T. M. Panggabean, J. Purba, A. Sutiani, and M. A. Panggabean. 2022. Analisis Hubungan Antara Kemampuan Matematika dan Analisis Kimia Terhadap Hasil Belajar Kimia Materi Kesetimbangan Kimia," *Jurnal Inovasi Pembelajaran Kimia. (Journal Innov. Chem. Educ)*, vol. 4, no. 1, pp. 18–30.

[3] A. Sutiani, M. Situmorang, and A. Silalahi. 2021. Implementation of an Inquiry Learning Model with Science Literacy to Improve Student Critical Thinking Skills,. *International Journal of Instruction.*, vol. 14, no. 2, pp. 117–138.

[4] W. W. W. Brata, C. Suriani, H. Simatupang, S. Siswanto, and F. T. M. Panggabean. 2020. Prospective Science Teachers' Learning Independency Level on Blended Learning. *Journal of Physics: Conference Series 1462*, pp. 1–5, doi: 10.1088/1742-6596/1462/1/012070.

[5] J. Purba, F. T. M. Panggabean, and A. Widarma. 2022.

Development of Online General Chemistry Teaching Materials Integrated with HOTS-Based Media Using the ADDIE Model. *International Journal of Computer Applications Technology and Research.*, vol. 11, no. 05, pp. 155–159, doi: 10.7753/IJCATR1105.1001.

[6] J. Sianturi and F. T. M. Panggabean. 2019. Implementasi Problem Based Learning (PBL) menggunakan Virtual dan Real Lab Ditinjau dari Gaya Belajar Untuk Meningkatkan Hasil Belajar Siswa. *Jurnal Inovasi Pembelajaran Kimia. (Journal Innov. Chem. Educ)*, vol. 1, no. 2, pp. 58–63.

[7] F. T. M. Panggabean and J. Purba. 2021. Pengembangan E-Modul Terintegrasi Media Berbasis Adobe Flash CS6 Untuk Meningkatkan Kemampuan Pemecahan Masalah Kimia Mahasiswa," *Jurnal Inovasi Pembelajaran Kimia. (Journal Innov. Chem. Educ)*, vol. 3, no. 2, pp. 116–122.

[8] H. Fives, W. Huebner, A. S. Birnbaum, and M. Nicolich. 2014. Developing a Measure of Scientific Literacy for Middle School Students. *Science. Education.*, vol. 98, no. 4, pp. 549–580, doi: 10.1002/sce.21115.

[9] F. T. M. Panggabean, P. M. Silitonga, and M. Sinaga. 2022. Development of CBT Integrated E-Module to Improve Student Literacy HOTS. *International Journal of Computer Applications Technology and Research*., vol. 11, no. 05, pp. 160–164, doi: 10.7753/IJCATR1105.1002.

[10] I. S. Jahro, A. Darmana, and A. Sutiani. 2021. Improving Students Science Process and Critical Thinking Skills Using Semi-Research Patterns Practicum. *JTK Jurnal Tadris Kimiya.*, vol. 6, no. 1, pp. 82–91.

[11] A. Wahyuni and E. Yusmaita. 2020. Perancangan Instrumen Tes Literasi Kimia Pada Materi Asam dan Basa. *Edukimia*, vol. 2, no. 3, pp. 106–111.

[12] K. F. Simamora. 2022. Kemampuan HOTS Siswa Melalui Model PjBL Ditinjau dari Kemampuan Literasi Kimia Siswa. *Jurnal Inovasi Pembelajaran Kimia. (Journal Innov. Chem. Educ).*, vol. 4, no. 1, pp. 55–65.

[13] A. Rusilowati, L. Kurniawati, S. E. Nugroho, and A. Wdiyatmoko. 2016. Developing an Instrument of Scientific Literacy Asessment on the Cycle Theme. *International Journal of Environmental & Science Education.*, vol. 11, no. 12, pp. 5718–5727.

[14] A. D. Safitri, A. Rusilowati, and Sunanro, 2015. Pengembangan Bahan Ajar IPA Terpadu Berbasis Literasi Sains Bertema Gejala Amal. *Unnes Physics Educational Journal.*, vol. 4, no. 2, pp. 32–40.

[15] T. E. Yulianti and A. Rusilowati. 2014. Analisis Buku Ajar Fisika SMA Kelas XI Berdasarkan Muatan Literasi Sains di Kabupaten Tegal. *Unnes Physics Educational Journal.*, vol. 3, no. 2, pp. 68–72.

[16] A. T. P. Retno, S. Saputro, and M. Ulda. 2017. Kajian Aspek Literasi Sains Pada Buku Ajar Kimia SMA Kelas XI di Kabupaten Brebes, in *Prosiding Seminar Nasional Pendidikan Sains (SNPS)*, pp. 112–123.

[17] S. Rahayu. 2017. Mengoptimalkan Aspek Literasi dalam Pembelajaran Kimia Abad 21, in *Prosiding Seminar Nasional Kimia UNY 2017*, pp. 1–16.

[18] S. A. Rodzalan and M. M. Saat. 2015. The Perception of Critical Thinking and Problem Solving Skill among Malaysian Undergraduate Students. *Procedia - Social and Behavioral Sciences 172*, pp. 725–732, doi: 10.1016/j.sbspro.2015.01.425.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 241-246, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1010

Wine Quality Classification Using Machine Learning Algorithms

Agbo Chijioke Benjamin
Master of Science (Computer Applications)
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed) University
Pune 411016, Maharashtra, India

Abstract: It has been long-established that wine making is an old craft that requires deep knowledge about the conditions and components that may be present in a wine. The need for quality control has always played a crucial role in the production of wines. Different regulatory agencies stipulate permissible production strategies of using some of the additives and processing agents. Assessing the wine quality using the usual traditional methods is not only tedious but also lack that level of consistency and reproducibility in production. Modern, through the machine learning algorithms it's more fitted to predict with the help of an automatic predictive system infused into a decision support system. In this paper, I have explored different machine learning models for classifying wine quality based on various metrics and components associated to wine quality, the ranking of the wine quality as well as investigation surrounding wine taste differing from another using machine learning models such as Naive Bayes algorithm, K-Nearest Neighbor algorithm, and Support Vector Machines algorithm.

Keywords: Machine Learning, Classification, Naive Bayes, K-Nearest Neighbors, Support Vector Machines.

1. INTRODUCTION

It has been long-established that wine making is an old craft that requires deep knowledge about the conditions and components that may be present in a wine. In all countries, wine consumption frequency has gone high during the pandemic, as a result, winery should consider alternatives to improving the quality of the wine at less cost. It is observed that most of the chemical components used in wine production are same for different wine based on the tests, and each quality of the chemical composition have varying level of concentration or impact for each type of wine. Therefore, the need to classify the quality of the wine for quality assurance is very apt. This case study aimed at predicting the quality of a wine from feature sets given as an input of a rating scale of 0-10 as an output. The dataset consists of Input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol downloaded from [11]. Focus on red wine, quality is being rated on the scale values of [3,4,5,6,7,8], the quality gets better as the scale value increases(i.e. 3 = lowest quality and 8 = highest quality). By implementing the supervised machine learning algorithm, we can easily classify the quality of the wine as good or bad using classification approach. Here, we focus on each class of the wine individually in order to successfully determine decision boundaries that can be fit for prediction if new data is supplied to the model.

2. LITERATURE REVIEW

K. R. Dahal et al. [1] has implemented the prediction of wine quality using Ridge Regression(RR), Support Vector Machines(SVM), Gradient Boosting Regressor(GBR), and Multi-layer perceptron(MLP), according to the researchers, the evaluation of the result was performed with help of Mean Squared Error(MSE), Correlation Coefficient(R), and Mean Absolute Percentage Error(MPE). The result outputted from the model performance measurement metrics shows that Gradient Boosting Regressor(GBR) outperformed other three model on the Test dataset with MSE=0.3741, R=0.6057, and MPE=0.0873.

[2] This research aimed at detailed comparison and evaluation of wine quality between red wine and white wine with an inclusion of a grid search algorithm for model Accuracy improvement. Support Vector Machines(SVM), Naive Bayes, and Artificial Neural Network(ANN) was the machine learning algorithms used. The model performance measurement was evaluated using Pearson Coefficient Correlation(R), Accuracy Score, Precision Score, Recall Score, and F1 Score and was found that Artificial Neural Network(ANN) performs better than the other two models. The performance measurement metrics based on the accuracy scores are as follows:

Accuracy Score for Naive Bayes Algorithm for red wine is 46.33% and 46.68% for white wine, Accuracy Score for Support Vector Machines(SVM) for red wine is 83.52% and 86.86% for white wine, and Accuracy Score for Artificial Neural Network(ANN) for red wine is 85.16% and 88.28% for white wine.

According to a research work [3],the more fermentation yeast yields have an impact in maintaining the quality of the wine. In their paper, K-Nearest Neighbor(KNN), Support Vector Machines(SVM), J48(Decision Tree Algorithm), Random Forest Algorithm, CART(Decision Tree Algorithm), and MP5(Multiple Regression Model)was used to analyse red wine and white wine quality based on the model's performance measurement metrics, the accuracy scores of both was compared. The result showed that MP5(Multiple Regression Model) outperformed the other Models used in this research. The Accuracy scores result is shown below:

Accuracy Score for KNN Model for red wine is approx. 61% and approx. 61% for white wine, Accuracy Score for SVM Model for red wine is approx. 62% and approx. 64% for white wine, Accuracy Score for J48 Model for red wine is approx. 56% and approx. 69% for white wine, Accuracy Score for Random Forest Model for red wine is approx. 73% and approx. 76% for white wine, Accuracy Score for CART Model for red wine is approx. 71% and approx. 78% for white wine, and Accuracy Score for MP5 Model for red wine is approx. 82% and approx. 83% for white wine.

[4] tried to implement a feature selection technique that can be used to analyse the impact of the scientific tests. based on the result of the research, it has clearly demonstrated that not all the input characteristics has an impact on the quality. For instance, an increase in quality will not rapidly change the residual sugar level. The researcher's accuracy scores based on the four models implemented are random forest is 88%., Stochastic gradient descent is 81%, SVM is 85%, and Logistic Regression is 86%.

[5] The researchers focused on quality ranking and reason behind choice of wine taste for different people using an ML algorithm. From their studies, the quality of a wine is hugely dependent on the level of acidity present in the wine, lower the level of acidity, the higher the wine quality becomes. while volatile acidity indicates presence of unpleasant fragrance(bad quality). The model and performance metrics measurement was done using their accuracy scores: Logistic Regression of 86% accuracy scores, Stochastic Gradient Descent of 81% accuracy scores, SVM of 85% accuracy scores, and Random Forest of 87.33% accuracy scores.

[6] The author has implemented three regression techniques with SVM performing better than the multiple regression and artificial neural network methods. the model targets oenologist wine tasting analysis as well as wine production improvement.

[7] Emphasized two major approaches for wine quality prediction. first, the generalized approach, an algorithm that focuses on the implementation of hybrid model and Second, the genetic approach, an algorithm that tends to generate new offspring.

According to [8] the research was implemented using three different machine learning classification algorithms: Decision tree, Adaptive Boosting, and Random Forest, after applying performance measurement metrics Random Forest seems to perform better as compared to the other two models mentioned.

[9] The study focused on the correlation between sensory, volatile and elemental profiles of a wine to their quality proxies. They suggested that initial look at quality correlations is vital parameters for ensuring that a wine is of good quality.

[10] The researchers focused on the problems that needs to be solved in quality control in sensory method of expert testers and evaluations in wine manufacturing industries.

3. PROBLEM FORMULATION

Based on several research papers reviewed in this paper, from the performance measurement metrics most reported an accuracy score below 89%. Therefore, the focus of this project consists of two problems as explained below:

(1) A look at the significance features for making a prediction of wine quality.

(2) An improved performance of the prediction model using the three classifiers mentioned above.

The problem has carefully been addressed by implementing the following:

- ✓ The Need to balance the imbalance dataset
- ✓ The impact of the features needs to be analysed.
- ✓ Applying hyperparameter tuning to optimize the model classifier
- ✓ Finally, building the model and Evaluation of the processes.

4. METHODOLOGY

The main reason behind this research is to predict wine quality based on various metrics and Physicochemical properties associated to the wine, why people prefer a particular wine taste from another using machine learning models like Naive Bayes algorithm, K-Nearest Neighbor algorithm, and Support Vector Machines algorithm and then compare the result of the three classifiers to determine a model that perform better among them.

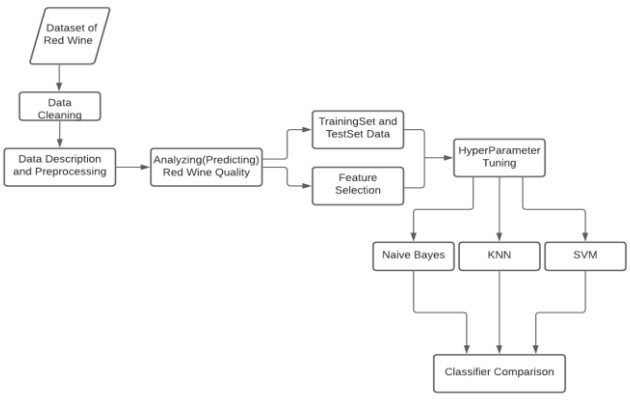


Figure 1: Red Wine Quality Prediction Model

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 241-246, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1010

5. DATASET AND EXPERIMENT DISCUSSION

The dataset used for this study was obtained from Kaggle repository and consists of 11 input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and 1 output variable: quality with rating scale of 3 to 8 (3 represent bad quality and 8 represent best quality). Python libraries used in the study are numpy, pandas, matplotlib, seaborn, imblearn and sklearn.

Experiment Steps

i. First, we start by setting up our environment and loading important libraries like numpy and pandas, also matplotlib and seaborn for the data visualization.
ii. After that, use the read_csv() method to read the data
iii. I have applied data cleaning method to check for duplicate records in the dataset.
iv. Then, check for null values present in the dataset.
v. In step 5, I have compared the correlation between quality and other composition, using corr() method on the quality column. we observed that alcohol and sulphates is highly proportional to quality and volatile acidity is inversely proportional to the quality.
vi. To check whether we have any pair of highly correlated independent features (Multicollinearity problems), we graphically represent the correlation with heatmap.
vii. Basically, six quality rating scale is used. here, we tried to split the wine quality column into two groups (0 and 1): [3,4,5,6] represent low quality wine and 0 is assigned to it, [7,8] represent high quality wine and 1 is assigned to it.
viii. We now check if the two classes(0 and 1) are balanced or not by visualizing them via pie chart.
ix. We conducted exploratory Data Analysis(EDA) by analysing features columns with histogram and boxplot -By using histogram(histplot()) we perform univariate analysis whereas boxplot(boxplot()) performs bi-variate analysis.
x. Creating a barplot to analyse each of the columns with quality column.
xi. Balancing the data point for each class using SMOTE technique.
xii. After up sampling, we check if the classes are balanced or not by plotting a pie chart.
xiii. It is a good practice to split the data into Training Set and Test Set to avoid Data Leakage. With Training Set having 75% of the dataset and Test Set with 25% of the dataset.
xiv. After balancing the data point, we then check if the columns still have skew by plotting the histogram of the training set.
xv. The skewed columns were fixed by applying power transformer on the selected columns(features).
xvi. Then applying the Feature Scaling to all the features with MinMaxScaler().
xvii. Testing and validating three types of classifiers: Naive Bayes, K-Nearest Neighbor, and Support Vector algorithm.
xviii. Evaluating the Performance measurement metrics of the models for comparing the three types of classifiers.

6. RESULT ANALYSIS

The analysed Red wine quality scale implementation was done in python programming environment using jupyter Notebook. The target(output) attribute in the dataset is quality column with scale ranging from 3 to 8, 3 being lowest quality and 8 being the highest quality. The analysis as represented in visuals and table are given below:

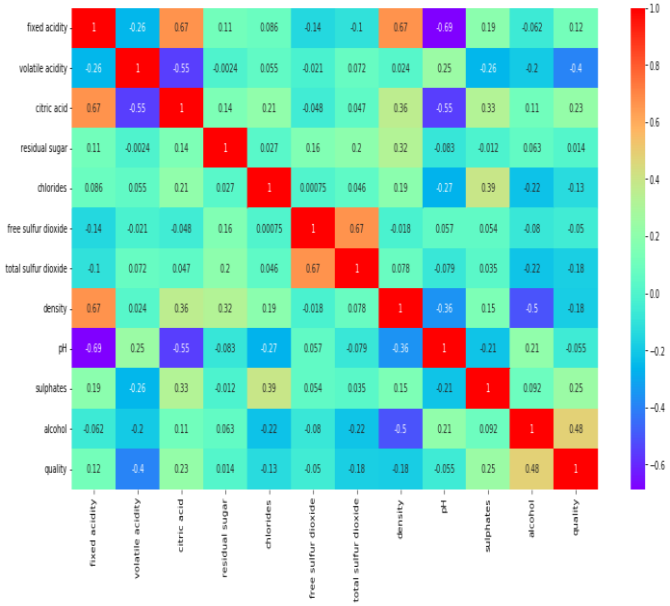


Figure 2: Correlation Matrix

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 241-246, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1010

Figure 2 represent the rank of features for the correlation matrix according to the high correlation values to the quality class such features are 'alcohol', 'volatile acidity', 'sulphates', 'citric acid', 'total sulfur dioxide', 'density', 'chlorides', 'fixed acidity', 'pH', 'free sulfur dioxide', 'residual sugar'.

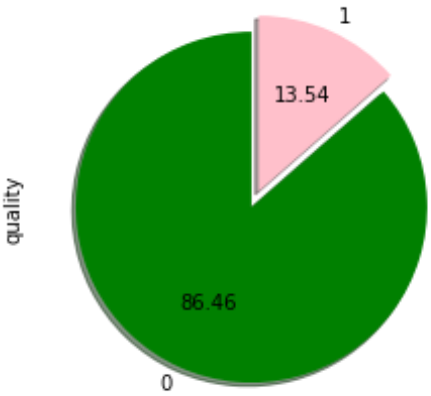


Figure 3: Imbalanced Class

Figure 3 shows an imbalance classes between the low-quality red wine represented by 0(86.46%) and high-quality red wine represented by 1(13.54%)

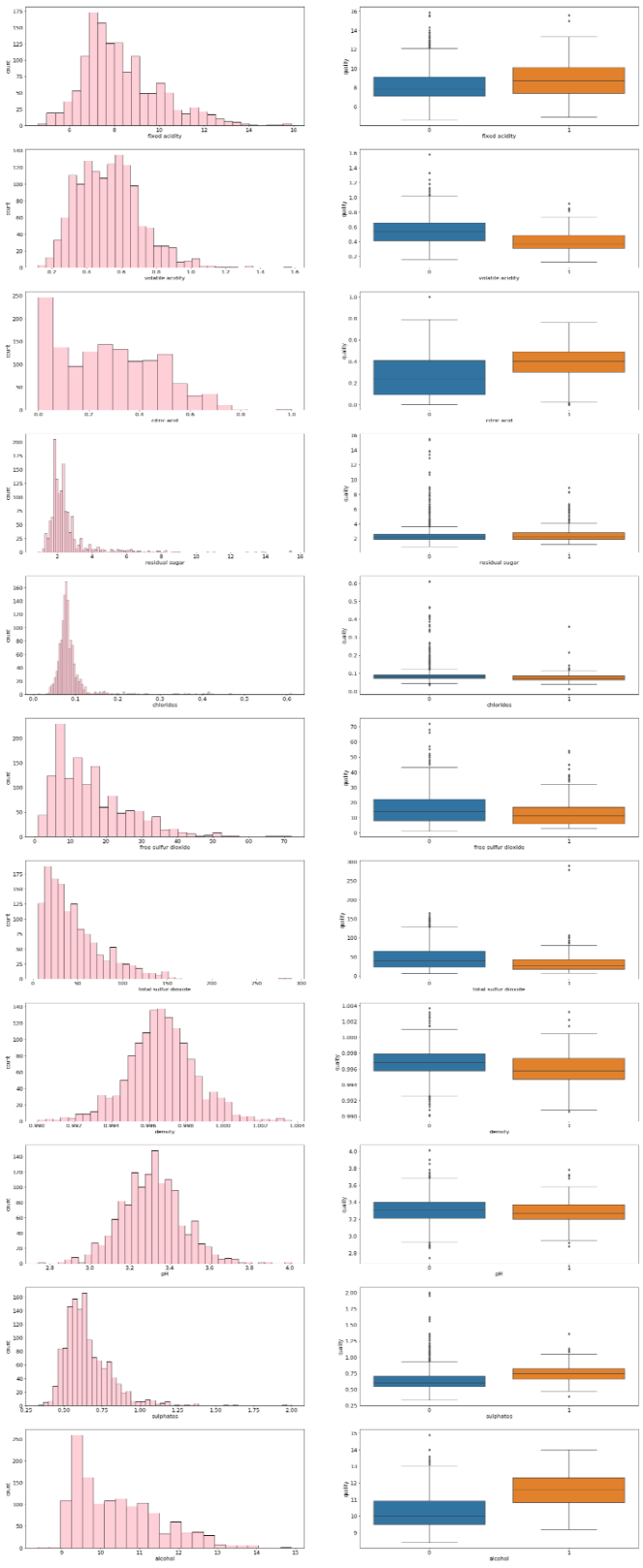


Figure 4: Analyzing each feature column with quality

Figure 4 shows presence of skewness in our feature columns in the histogram univariate analysis and the boxplot bi-variate analysis.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 241-246, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1010

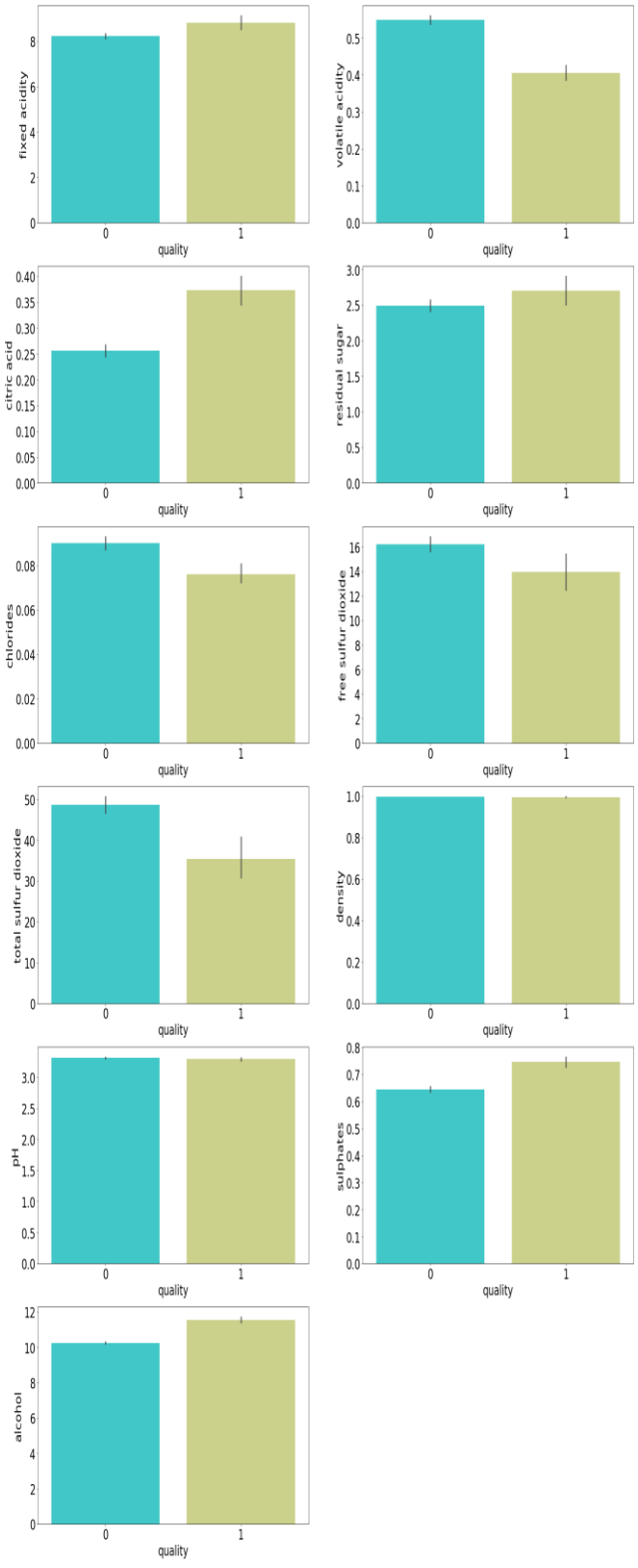


Figure 5: Analyzing each of the feature columns with quality using barplot

In figure 5, each feature column with respect to quality are represented in the bar plot by analysing low-quality and high-quality separately in order to determine the influence of individual feature on quality.

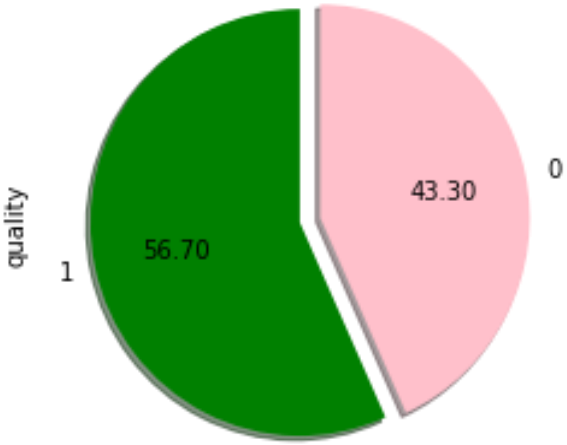


Figure 6: balanced Class

Figure 6 shows a refined or balanced class between the low-quality red wine represented by 0(43.30%) and high-quality red wine represented by 1(56.70%)

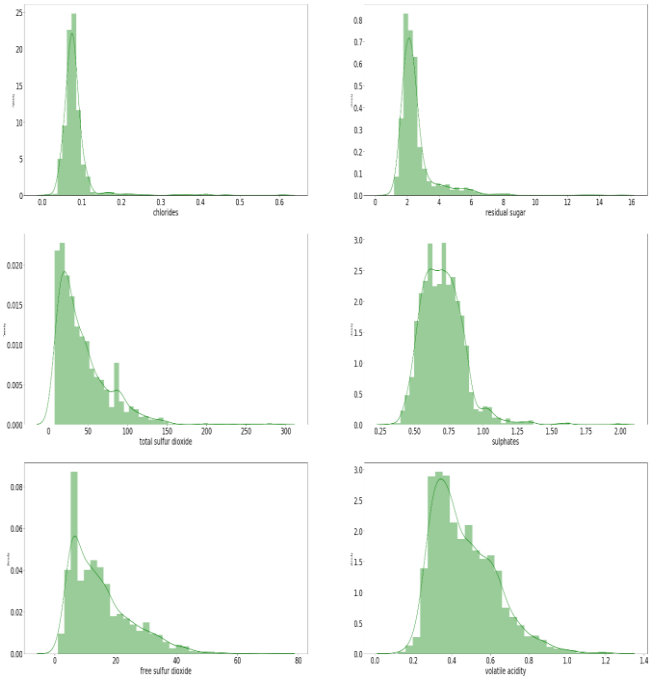


Figure 7: Histogram representing the skewed columns

From figure 7, we have represented six feature columns perfectly skewed on the Training Set.

Algorithm	Accuracy Score	Precision Score	Recall Score	F1 Score
Naïve Bayes	89	89	91	90
K-Nearest Neighbor	95	92	99	96
Support Vector Machines	96	95	97	96

Table: 1

Table 1 illustrates the overall performance measurement metrics results obtained from the study in this project.

7. DISCUSSION

In this study, the algorithms we have implemented for the classification are:

i. Naive Bayes Algorithm
ii. K-Nearest Neighbor (KNN) Algorithm
iii. Support Vector Machine (SVM) Algorithm

After analyzing each classifier, we then compared results predicted from the three algorithms as follows: Naive Bayes Algorithm has given accuracy of 89%, K-Nearest Neighbor Algorithm with an accuracy score of 95%, and Support Vector Classifier has given an accuracy score of 96%. Based on the accuracy score results, it's clear that SVM outperformed both Naive Bayes and KNN algorithms.

8. CONCLUSION

From the bar plot in figure 5, we can agree that in as much as every feature column (Physicochemical properties associated to the wine) may impact on the wine quality but some may not affect the dataset, based on the bar plot on residual sugar column it is obvious that as the quality rises, the residual sugar remains normal. unlike volatile acidity, alcohol or citric acid column that shows drastic change with increase in quality.

9. FUTURE SCOPE

In future, we recommend implementing new performance measurement metrics as well as algorithms for more refined scores and better comparison. In so doing, wineries can predict the quality of different varieties of wine with a better accuracy, which in turn can enhance future product.

10. REFERENCES

1. Dahal, Keshab & Dahal, Jiba & Banjade, Huta & Gaire, Santosh. (2021). Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics. 11. 278-289. 10.4236/ojs.2021.112015.

2. Kothawade, R. D. (2021). Wine quality prediction model using machine learning techniques (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-20009

3. Gupta, Mohit & Chandrasekaran, Vanmathi. (2021). A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality. International Journal of Recent Technology and Engineering. 10. 314-321. 10.35940/ijrte.A5854.0510121.

4. Devika Pawar, Aakanksha Mahajan & Sachin Bhoithe. (2020). Wine Quality Prediction using Machine Learning Algorithms. International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 385-388, 2019, ISSN:-2319–8656

5. Sinha, Anurag & kumar, Atul. (2020). Wine Quality and Taste Classification Using Machine Learning Model. International Journal of Innovative Research in Applied Sciences and Engineering. 4. 715-721. 10.29027/IJIRASE.v4.i4.2020.715-721.

6. Nikita Sharma, "Quality Prediction of Red Wine based on Different Feature Sets Using Machine Learning Techniques", International Journal of Science and Research (IJSR), Volume 9 Issue 7, July 2020, 1358 - 1366

7. S. Aich, A. A. Al-Absi, K. L. Hui, J. T. Lee and M. Sain, "A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques," 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018, pp. 139-143, doi: 10.23919/ICACT.2018.8323674.

8. G. Hu, T. Xi, F. Mohammed and H. Miao, "Classification of wine quality with imbalanced data," 2016 IEEE International Conference on Industrial Technology (ICIT), 2016, pp. 1712-1217, doi: 10.1109/ICIT.2016.7475021.

9. Hopfer, Helene & Nelson, Jenny, Jennifer & Ebeler, Susan & Heymann, Hildegarde. (2015). Correlating Wine Quality Indicators to Chemical and Sensory Measurements. Molecules. 20. 8453-8483. 10.3390/molecules20058453.

10. McGrew, D. & Chambers, Edgar. (2012). Sensory quality control and assurance of alcoholic beverages through sensory evaluation. 10.1533/9780857095176.1.24.

11. https://www.kaggle.com/sgus1318/winedata.