
CLOUD MODEL CONSTRUCT FOR TRANSACTION-BASED COOPERATIVE SYSTEMS

A. A. Ajuwon¹ and R. O. Oladele²

¹Emmanuel Alayande College of Education, Oyo

²University of Ilorin, Ilorin

¹ayoajuwon@gmail.com, ²roladele@unilorin.edu.ng

ABSTRACT

Plethora of cooperative enterprise models exist in practice. Inherent complex and tasking operations and lack of financial strength to procure cutting edge Information Technology infrastructure are some of the problems faced by these cooperative enterprises. In this paper, a cloud model is constructed for transaction-based cooperative systems with a view to mitigating these problems. The model is implemented using a server script language (PHP) and Mysql database engine. Test results show that while database security issues can be intuitively tackled by using Mersenne Twister random number algorithm, server scripts are secured through simple abstraction. Results also reveal that credit cooperatives will benefit from the model by taking advantage of its low initial investment feature to make them part owner of the cloud software. It is therefore safe to conclude that while cooperative accounting is a new area for research attention, this model can solve the problems associated with cooperative administration.

Keywords: Cloud, Cooperative Enterprise, Database Engine, Transaction, Algorithm, Model.

1.0 INTRODUCTION

Cloud computing is a general term for anything that involves delivery of services that are hosted over the Internet. In cloud computing the end users' knowledge is neither necessary nor required. Also, it is not necessary to have the physical infrastructures in place to access the service. The software, database and other applications are delivered from web browsers and mobile applications. It allows the users to access the applications using the web browsers as if it is done on their own personal computers, mobile phones, or personal digital assistants. The services can be delivered to individuals, communities or to large corporate or government agencies. The services are provided based on the client's needs, so there is no need to make huge investment on procuring the hardware and software required to provide needed services or incurring unnecessary personnel and operation overheads.

The term cloud computing can be used in many ways. Some consider it to be a pool of virtualized computer resources and others say it is the dynamic development and deployment of

software fragments [1]. Macherla [2] referred to Cloud as a "model" and not as a "technology" per se that enables banks to use hardware and software via the internet as a paid service.

Cloud service models have been classified into three classes according to the abstraction level of the capabilities and resources provided and the service model of providers: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [3]. This work is particularly averred to software as a service (SaaS) model, where users purchase access and use an application that is hosted on the internet (the cloud). In this model, end user application is delivered as a Service, Platform and Infrastructure are abstracted, and can be deployed and managed by a third party with little efforts.

A transaction-based cooperative is a society owned by members who enjoy equal rights in the affairs of the society. The society avails its members the opportunity of engaging in periodic savings. It also encourages its members to patronize and transact business with the

cooperatives' owned commodity sales outlets. The society provides funds for its members in form of loans which they service at scheduled intervals. These societies can be owned by members of the same workplace such as Unilorin cooperatives, EACOED cooperatives, FUNAB cooperatives, and religious associations such as Muslim cooperatives or Christian cooperatives, etcetera.

In [4], it was pointed out that Cooperative Enterprise has a long history and that its origin dates back to the 15th Century. However, it was the establishment of the **Rochdale Society of Equitable Pioneers in 1844** that is viewed as the foundation of the modern cooperative movement. The so-called Rochdale Pioneers were ambitious and had lofty goals for their cooperative namely: (1) to sell provisions at the store; (2) to purchase homes for their members; (3) to manufacture goods which their members needed; (4) to provide employment for their members who are either out of work or are poorly paid [5].

Cooperative Enterprises started in many countries several years before policy formulation through Government gazette and exemptions from banks. In Nigeria, modern cooperative societies started as a result of the Nigerian cooperative society law enacted in 1935 following the report submitted by C. F. Strickland in 1934 to the then British colonial administration on the possibility of introducing cooperatives societies into Nigeria [6].

2.0 REVIEW OF RELATED WORK

Researchers at International Labour Organization reviewed different models of cooperative development and were able to identify the impact of liberalization measures on these models. The research which was carried out in 11 African countries revealed that cooperatives in Africa have survived the market forces and continued to grow in number and membership. They observed a slow but sure erosion of the unified model and the adoption of a social economy model [7].

In [8], the authors developed a dynamic system that effectively manages the loan scheme of a named organization. The system essentially manages both short-term and long-term loans, and keeps track of cash inflow and outflow of a cooperative society among others. It utilized

Structured Query Language (SQL) Server database architecture at the back end and Visual Basic.Net framework at the front end. This makes it user-friendly and highly interactive. The Object Modeling Technique (OMT) is adopted for the analysis and design of the Loan Software. They concluded that achieving optimal performance requires a means to both measure performance and plan sufficiently for the future so that the computing infrastructure will meet the demand of automation. However, their work did not adopt cloud computing technology.

Macherla [2].produced a white paper from the point of view of a banking software development organization. The article expressed the need for cloud in banking. It opined that leveraging technology to service a wide range of customers and achieving customer satisfaction is what determines banking success. The paper concluded that Technology is a double-edged sword, on one hand, it is inevitable and on the other hand, prone to obsolesce. The paper only described "Cloud Model" and its benefits, it never implemented it.

In [9], the author focused on the Co-operative Banking model and the Grameen Bank model. The study attempts to highlight their histories, institutional arrangements, the design of their saving and loan delivery systems and most importantly their strengths and weaknesses. The study emphasizes the use of smart card technology for the purpose of making easy withdrawals through the Automated Teller Machine (ATM), and controlling the withdrawals by placing a withdrawal limit if there is a loan taken out by the member, and experimenting with a complimentary currency.

In [10], the author wrotefor an in-house software development firm, Temenos Ren Money Nigeria Ltd. Her paper introduced a microfinance software solution that would allow instant access for both core banking and risk management platforms. The software firm included cloud technology features in its core banking module. It is however a core banking software and not a cooperative software.

In [11], the authors studied The Role of Credit Co-Operatives in the Agricultural Development of Andhra Pradesh, India. The main objective of

the study was to evaluate the performance of Credit Cooperatives by analyzing its deposits, credit and impact of credit on the beneficiaries. Through cooperative credit, the farmers benefitted to a maximum extent by increasing their agricultural output which in turn increased their levels of employment and income. It was concluded that cooperative credit has become a powerful tool in the agricultural development of the state. In this study no particular reference was made to the role of information and communication technology on the cooperative enterprise.

Even though many cooperative societies have set up websites that give information about their history, products, and services, and have global presence on the internet, those that have enjoyed the benefits provided by cloud technology are still very scanty. Today, cloud computing is reshaping how businesses are done, managed and transformed at lower cost anywhere and at any time. To a large extent, this transformation is being enforced by the liberalization and globalization of markets and the growing use of cloud computing technologies.

3.0 METHODOLOGY

Exploratory Research Method is adopted for the initial information gathering and **Object-Oriented Technology** is adopted for Analysis, Design, and Implementation. The Unified Modeling Language (UML) is used to construct Use Case diagrams, Class diagrams, Sequence diagrams, and activity diagrams. Software as a Service model of the cloud technology where platforms and infrastructure facilities are abstracted forms the basis of the model construction. The model has four basic components: The database, Scripts, The client application, and Object oriented analysis and design.

Database: A database is created for every cooperative society that subscribes to the software in the cloud. The database contains the basic information about the cooperative society. Each database consists of tables containing

information about the society members and details of their loans, commodity, and investments transactions. **MYSQL** Database Engine is adopted for the software. This is because it is easy to obtain (free versions are available on the internet) and it is very stable.

Scripts: The scripts act as middleware between the database and the client application. These scripts are developed to update the database, generate user reports, and carry out formatting and security checks on the input data. **PHP** is adopted as scripting language for the software. This is because it is server based and can be abstracted.

Client Application: This application provides the interface between the database and the users. **Opera Mini** is adopted as the default web browser application for interacting with the system. This is because it has an embedded java scripting language and is available on most mobile devices. Many other popular browsers (i.e. Internet Explorer, Mozilla Firefox) require the subscriber to install java scripting language on their infrastructure.

Object oriented analysis and design: The goal of the object-oriented analysis is to understand the domain of the problem and the system's functionalities. This is achieved by identifying the users, usually called the Actors, Developing the Use Case scenarios, Developing interaction diagrams, and Identifying the classes and model the system using the Class Diagrams.

3.1 USE CASE DIAGRAMS

The Use Case diagrams are considered for high level requirement analysis of the system. After the requirements of the system are analyzed, the functionalities are expressed as use cases. Here, the actors or users are identified as the cooperative management, cooperative assistants, members of the cooperative society, and the Web 2.0 System. The use cases are identified and the relationship between the actors and the use cases are as illustrated in figure 1 below.

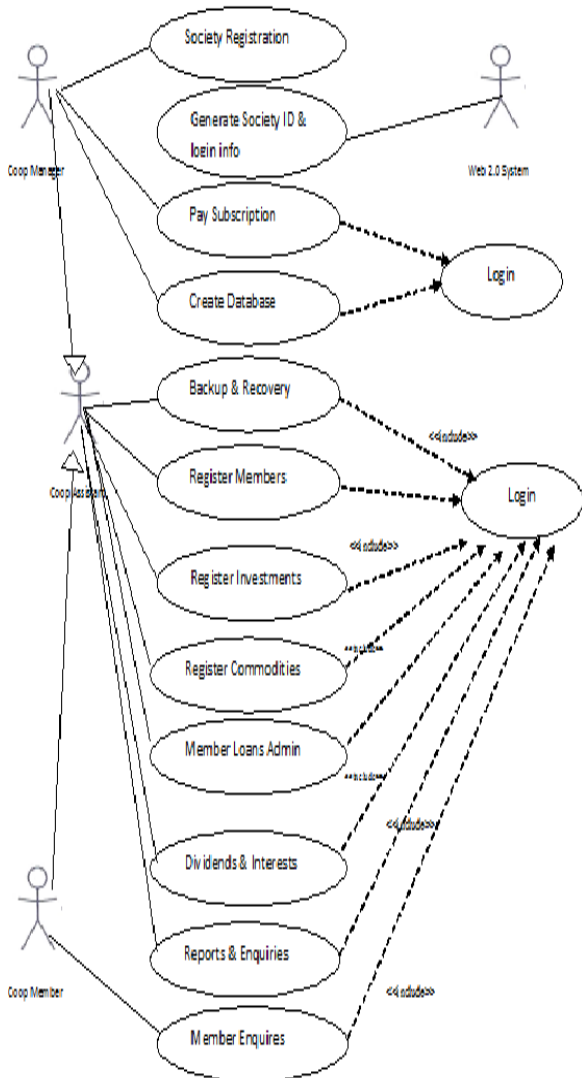


Figure1 Use Case Diagrams

3.2 UML CLASS DIAGRAMS

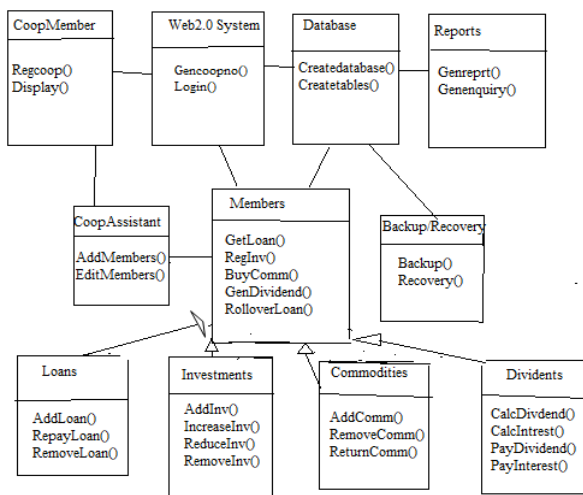


Figure 2: Class Diagrams

The UML class diagram models the static structure that represents the fundamental architecture of the system. From the description of the Use Case Diagrams the objects involved in the system are identified as the Web 2.0 system, Database, Coop manager, Coop assistant, Coop members, Loans, Investments, Commodities, Dividends, Backup/Recovery system, and Reports. These objects are specified with the services which they provide in the above UML class diagram. The Class Diagram shows the classes and how they are related to each other.

3.3 SEQUENCE DIAGRAMS

The sequence diagram is used to model the flow of control between objects. It shows how the objects interact dynamically over time and how messages are passed between them. In this study the flow of control between the objects are modeled thus:

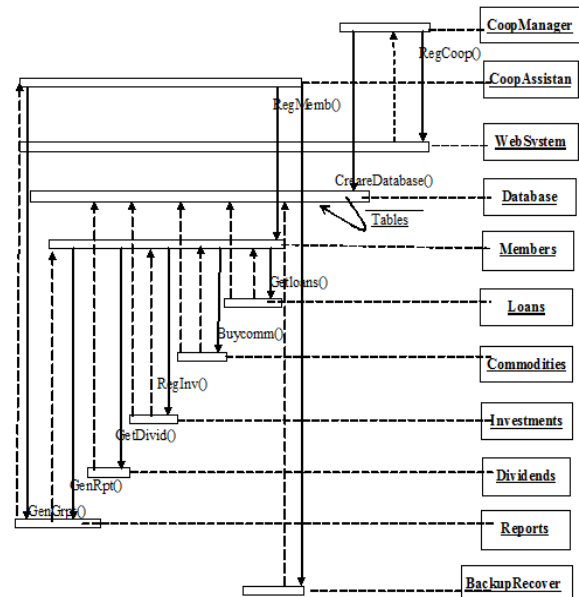


Figure 3 Sequence Diagrams

3.4 ACTIVITY DIAGRAMS

The dynamic nature of this system can also be modeled with an activity diagram. Here the focus is on the activities and responsibilities of the objects that drive the system without showing the message flow. The diagram exhibits capabilities such as parallel flows, concurrency, and swim lanes. The four categories of users identified are: Coop Manager, Coop Assistant, Coop Members, and Web 2.0 System. Each user is represented as a swim lane in the model and

their activities are modeled in the following diagrams:

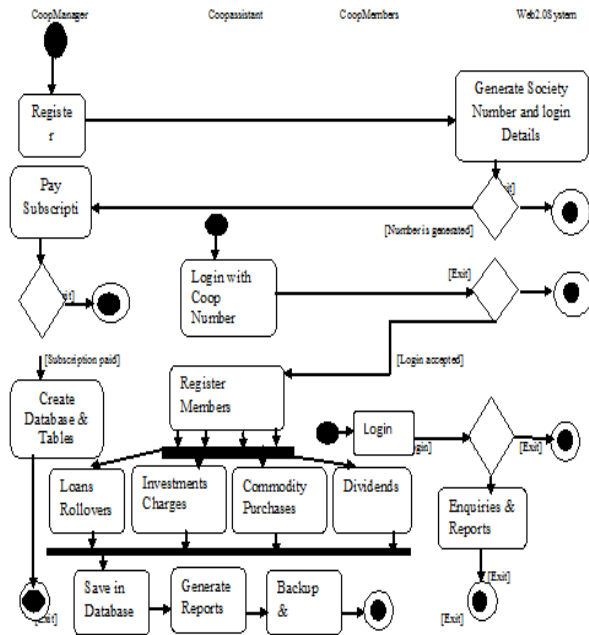


Figure 4 Activity Diagrams

3.5IMPLEMENTATION DIAGRAMS

Component diagrams are used to describe the implementation view of a system. It models the physical artifacts such as the executables, libraries, documents, and files that reside in the node. The relevant artifacts and their relationships are as shown in the figure 5 below

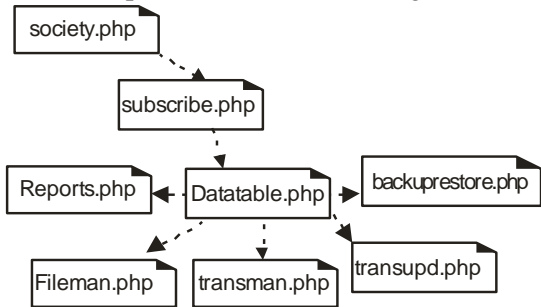


Figure 5: Components Diagrams

3.6DEPLOYMENT DIAGRAMS

Deployment diagrams are used for describing the hardware components on which software components are deployed at run time. The diagram is also used to show the topology of the hardware components that are deployed. This diagram however, modeled how the software is deployed.

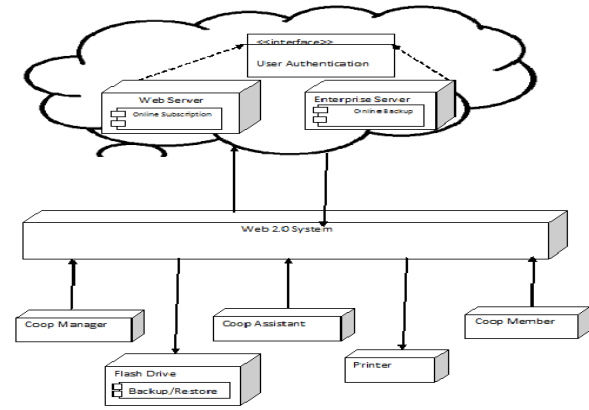


Figure 6: Deployment Diagrams

4.0 RESULTS AND DISCUSSIONS

4.1 RESULTS

After the subscriber has registered his cooperative society, he proceeds to pay his annual subscription, and then continues to enter the required information into the cloud, and later generates necessary reports. This is achieved using the following web pages:



Main Menu

Register Society
Pay Subscription
Create Your Database
Log into your Database



Commodity Types
Commodity Updates
Society Update

Standing data

Membership Register
Loan Register
Investment Register

Cloud Model Construct For Transaction-Based Cooperative Systems
A. A. Ajuwon and R. O. Oladele

Registers



COOPERATIVES
 CLOUD MODEL

Home Standard Data Registers Transactions Reports BackupRestore Logout

Commodities					
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">Sales</td></tr> <tr><td style="padding: 2px;">Purchases</td></tr> </table>	Sales	Purchases			
Sales					
Purchases					
Loan Repayment					
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">Regular repayment</td></tr> <tr><td style="padding: 2px;">Regular Repayment Update</td></tr> <tr><td style="padding: 2px;">Bulk Loan Repayment</td></tr> <tr><td style="padding: 2px;">Loan Rollover</td></tr> </table>	Regular repayment	Regular Repayment Update	Bulk Loan Repayment	Loan Rollover	
Regular repayment					
Regular Repayment Update					
Bulk Loan Repayment					
Loan Rollover					
Investment					
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">Regular Contribution</td></tr> <tr><td style="padding: 2px;">Regular Repayment Update</td></tr> <tr><td style="padding: 2px;">Bulk Contribution</td></tr> <tr><td style="padding: 2px;">Bulk Contribution Update</td></tr> <tr><td style="padding: 2px;">Monthly Transaction Update</td></tr> </table>	Regular Contribution	Regular Repayment Update	Bulk Contribution	Bulk Contribution Update	Monthly Transaction Update
Regular Contribution					
Regular Repayment Update					
Bulk Contribution					
Bulk Contribution Update					
Monthly Transaction Update					



COOPERATIVES
 CLOUD MODEL

Home Standard Data Registers Transactions Reports BackupRestore Logout

Commodities							
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">Commodities List</td></tr> <tr><td style="padding: 2px;">Sales List</td></tr> <tr><td style="padding: 2px;">Purchases List</td></tr> </table>	Commodities List	Sales List	Purchases List				
Commodities List							
Sales List							
Purchases List							
Registers							
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">Loan List</td></tr> <tr><td style="padding: 2px;">Investment List</td></tr> <tr><td style="padding: 2px;">Membership List</td></tr> <tr><td style="padding: 2px;">Members Individual Detail</td></tr> <tr><td style="padding: 2px;">Summary of monthly payments</td></tr> <tr><td style="padding: 2px;">Detail monthly Payments</td></tr> <tr><td style="padding: 2px;">Summary of bulk payments</td></tr> </table>	Loan List	Investment List	Membership List	Members Individual Detail	Summary of monthly payments	Detail monthly Payments	Summary of bulk payments
Loan List							
Investment List							
Membership List							
Members Individual Detail							
Summary of monthly payments							
Detail monthly Payments							
Summary of bulk payments							
Loan Exceptions							
Dividend Reports							

Transactions

Reports

SAMPLE REPORTS

LIST OF REGISTERED PRODUCTS AS AT 14-08-2015

Product Name	Product Code	Ctr Code	Stock Level	Reorder Level	Last Order date	Loan Type	Pay No
deep prizer 350	1001	0	2	1		Essential	7

xtian coop 4


LIST OF REGISTERED LOANS AS AT 14-08-2015

Last Name	First Name	M. Number	Loan Type	Face Value	Rate	Loan Balance	Last Payment Date
ADEGBIJI	SAMUEL	300	Casual	30000.00	0.00	30000.00	2015-08-13


xtian coop 4

LIST OF REGISTERED INVESTMENTS AS AT 14-08-2015

Last Name	First Name	M. Number	Investment Type	Total Investment	Rate	Last Payment Date
ADEGBIJI	SAMUEL	300	Esther	7000.00		2015-08-14
ADEGBIJI	SAMUEL	300	Savings	5000.00		2015-08-13
ADEGBIJI	SAMUEL	300	Shares	5300.00		2015-07-31
ADEGBIJI	SAMUEL	300	Xmas	2000.00		0000-00-00

Last Name	First Name	M. Number	Cell Phone No	E-Mail	Sex	M. Status	Picture
ADEGBIJI	SAMUEL	300			Male	Married	

MEMBER DETAILS AS AT 14-08-2015

	300	ADEGBIJI	SAMUEL
---	-----	----------	--------

Type	Balance	Int. Rate	Paid Installment	Last Pay Date
Casual	30000.00	0.00 0	0	2015-08-13
Esther	7000.00		0	2015-08-14
Savings	5000.00		0	2015-08-13
Shares	5300.00		1	2015-07-31
Xmas	2000.00		0	0000-00-00


xtian coop 4

SUMMARY OF MONTHLY PAYMENTTS AS AT 14-08-2015

M. NUMBER	LAST NAME	FIRST NAME	LOANS	INVESTMENT	TOTAL
300	ADEGBIJI	SAMUEL	5000.00	300.00	5300

xtian coop 4

MEMBER MONTHLY PAYMENT DETAILS AS AT 14-08-2015

	300	ADEGBIJI SAMUEL ADEOLA	LOANS = 5000.00	CONTRIBUTIONS = 300.00	TOTAL = 5300
---	-----	---------------------------	--------------------	---------------------------	-----------------

DETAILS

Type	Description	Amount	Paid Installment	Last Pay Date
Casual	Loan Repayment	5000.00	6	0000-00-00
Shares	Shares	300.00	1	2015-07-31

xtian coop 4

MEMBER MONTHLY BULK PAYMENT DETAILS AS AT 14-08-2015

300	ADEGBIJI	SAMUEL	ADEOLA		6000.00	6000
-----	----------	--------	--------	--	---------	------

xtian coop 4

LOANS EXEPTION REPORT AS AT 14-08-2015

M. Number	Names	Loan Type	Loan Balance	Monthly Payment	Last Payment Date
300	ADEGBIJI SAMUEL ADEOLA	Casual	30000.00	5000.00	2015-08-13

4.2 DISCUSSIONS

Security Issues

In cloud computing, security issues have always been a topic of debate. As such, security is an important aspect of the design process in this work. Security has been handled from two levels: database and the script.

Database Protection

The first idea was to protect the database from intruders. So immediately the subscriber registers the society the user id and password are generated using Mersenne Twister random number algorithm. This algorithm also generates the database name. A good feature of the algorithm is that the number generated is a mixture of digits and characters which makes it more difficult to guess. It is also possible to specify the number of characters that is needed for any purpose. Even though the software recognizes the database of each subscriber as he logs into his database, the database is abstracted.

Again, the database is further protected by allowing the subscriber to have a complete backup of his database. This guides against any unforeseen disaster and gives the subscriber the assurance that his database is save anytime.

Script Protection

The second idea was to protect the script from malicious hackers. In order to achieve this, the frontend Hypertext Markup Language (HTML) is left as open source and the actual scripts were intentionally kept away from the browser. This action will prevent any malicious hackers from having access to the scripts thereby further keeping away the database from intruders.

5.0 CONCLUSION

A cooperative cloud model has been constructed using Object-oriented approach and was successfully implemented with PHP script and MySql database engine. The model software has been recommended for use as a service on the internet (Cloud).

Again, the subscribers will benefit from this research once the software model is published on the cloud, the subscriber does not need huge investment for him to use the software on the cloud. He only needs to pay a token to be able to access the software,

Moreover, the subscriber's information is very secure, because the security design concept of the software model has been treated from two levels: database level and script level. In addition, the subscriber is allowed to have a complete backup of his database which he can copy after using the software or restore before using it.

Cooperative accounting is another area that can be studied. Different cooperative accounting systems are used across the globe which can also be harmonized and standardized, so a research effort in this area is worthwhile. The constructed cloud model covers cooperative administration and reporting, this can be expanded to accommodate cooperative general ledger accounting. The combined model if implemented will improve the robustness of the system.

It is believed that there can be no information technology if there are no new software developments to manage new areas of human endeavors. Hence, the construction of this model and its implementation is an attempt to address new areas of information technology management. In addition, this model has proved that software can be used as a service for a transaction-based cooperative administration.

Again, the implementation of this model has shown that the operations of cooperative societies across the globe can be harmonized and standardized. The model has been able to accommodate differences in information requirements across the globe. This is because differed cooperative models across the globe have been carefully reviewed and merged to arrive at the final construction of the model.

This models is recommended for use by a cooperative society for a period of six months. During this period, its correctness should be monitored, scrutinized, and any discrepancies should be identified. At the end of the testing period the model can be deployed and released for public use.

REFERENCES

- [1] A. Anandasivam, B. Blau, J. Stosser and C. Weinhardt. "Business Models in the Service World." *IT Professional*. Vol 11 No. 2 pp28-33 Sept 2009.
- [2] N. B. Macherla. "Leveraging cloud in a competitive banking environment." www.infosvs/finac/c/solutions/toimht-papers/document/lcveramiim-cloLid-competitive-banking-envivironment.pdf, April 9, 2014.
- [3] J. Broberg, R. Buyya and M. Goscinski. *Cloud Computing: Principles and Paradigms, USA*: Wiley Press, 2011.
- [4] T. Mazzarol. *Cooperative Enterprise: A Discussion Paper and Literature Review*. University of Western Australia. 2009. pp8.
- [5] A. Z. Kimberly and C. Robert. *Cooperatives: Principles and Practices in the 21st century*. Cooperative Extension Publishing, Midson. 2004. pp12.
- [6] E. Onouha. "A critique of the drafts of cooperative policy for Nigeria." *Nigeria Journal of Cooperative Studies*. Vol.2No 1. 2002.
- [7] O.W. Fredrick, D. Patrick and P. Ignace. "Reinventing the wheel? African cooperatives in a liberalized economic environment." *Annals of Public and Cooperative Economics*, Vol. 80 No. 3. Sept 2009.
- [8] B.C.E .Mbam and K.O Igboji.. "Enhancing Cooperative Loan Scheme Through Automated Loan Management System." *West African Journal of Industrial & Academic Research*. Volt 6 No. 1 March 2013
- [9] V. Satgar. "Comparative Study- Cooperative Banks and the Grameen Bank Model. *Co-operative and Policy Alternative Center*." www.coiKic.or.za/tiles/cooverative-bank-and-the-Grameen-bank-model.pdf, May 5, 2014.
- [10] A. Lody. "Temenos Case study: A fully cloud-based microfinance software." www.temenos.com/documents/mi/cs/temenos_renmonv_cs.pdf. April 20, 2014.
- [11] R. Devi and S.R.K. Govt. "The Role of Credit Co-Operatives in the Agricultural Development of Andhra Pradesh, India." *International Journal of Cooperative Studies*. Vol. 1 No. 2, pp 55-64 2012.

A PROACTIVE APPROACH OF MEASURING THE
IMPACTS OF INFORMATION SECURITY AWARENESS ON
SOCIAL NETWORKS

J. O. K. Okesola¹, A. A. Owoade², A. S. Ogunbanwo³

¹Covenant University, Ota, Nigeria

^{2,3}Tai Solarin University of Education, Ijebu-Ode, Nigeria.

¹48948535@mylife.unisa.ac.za; ²owoadeakeem@yahoo.com; ³ogunbanwoafolakemi@yahoo.com

ABSTRACT

Nowadays, many measurement techniques are being implemented to determine the effectiveness of security awareness on social networks (SN). While these techniques are inexpensive, they are all incident-driven as they are based on the occurrence of incidence or success of attack(s). Additionally, they do not present a true reflection of awareness since cyber-incidents are hardly reported. The techniques are therefore adjudged to be post-mortem and risk permissive, the limitations that are unacceptable in industries where incident tolerance level is very low. This study deploys a password cracker technology, as a *non-incident statistics approach*, to proactively measure the impacts of awareness on SNs.

Key words: Security awareness; measurement techniques; non-incident statistics approach; password cracker; social networks; Socialist Online.

1. INTRODUCTION

Many experts agree that information security awareness (awareness) is effective while some others are of a different opinion [28]. Okesola in [21] particularly criticises the impacts of awareness, arguing that awareness has been promoted for many years as being fundamental to Information systems (IS) practices. He confirms that only very few studies have been done regarding its effectiveness and efficiency.

The importance of security awareness is actually being discussed by many authors and organisations but very few empirical studies are done, and none of these offers a technique that is effective in measuring users' behaviour in Social Network Sites (SNs) [28]. Similarly, very few experiments are done in the *measurement* of the effectiveness of the changes in human behaviour or attitude [26, 29]. Although recent research works [2; 30] are already looking into the impacts of awareness on SNs, they have always been focusing on the effectiveness of phishing tests, class-room based training, e-mail based training and web-based awareness material [13].

Research has been exhausted in the realm of awareness, but literature still lacks proof of the effectiveness of awareness methods from psychological theories and they are still silent on the fundamental assumptions of these methods. However, in their own study concluded in March 2011, Khan, Alghathbar and Khan evaluated the effectiveness of different awareness tools and techniques on the basis of psychological models and theories [14]. They succeeded in describing processes needed to measure awareness in an organisation.

In a report [1] concluded in November 2012, ABC classified all these methods as *incident statistics approaches* to measure awareness. The methods are adjudged to be incident statistics, since the measurement of their effectiveness is based on the occurrence of an event or success of an attack. Therefore, putting in mind the goal of password cracker as a non-incident statistics technique, this author chooses to develop a strong password cracker capable of cracking SN passwords that is hashed with a more complex system.

People generally do not see legitimate reasons behind the creation of password cracker. However, the problem is not the existence of password crackers, but their frequent illegal use by fraudulent people for bad goals and objectives. When employed with good intentions, password crackers offer a valuable service to system and data administrators by alerting them of system or users' weak passwords [27].

A review of the existing crackers and analysers did not provide any suitable cracker to capture the logon details of users on socialist Online, which includes passwords and user-Id [1]. Although many researchers have developed numerous crackers to crack and analyse user passwords of SNs, none of these solutions have ever worked with applications that do not use MD5 for password hashing [22]. Therefore, the author had to develop a new password cracker suitable for this research purpose.

This study begins with a comprehensive literature review to highlight the main definitions, theories, models and empirical findings that provide background to the research in question. The literature overview reviews the existing related works on awareness as related to SNs and provides the theoretical foundation for the research questions and sub-questions. It is on this basis that the initial study is planned and delimited.

2. RELATED WORK

While several literature sources emphasise the importance of awareness in SNs, some notable authors in [3],[7],[11],[23], and [24] surprisingly argue that security education may yield negative results against the expectation and thereby promote the potential risks that users are exposing themselves to. They report that if users are well informed about the risk in disclosing credit card details via emails, but the attack approach is changed to be launched through telephone requests, then such users could be at risk for simply following what they were told to do.

Research has been exhausted in the realm of awareness but literature still lacks proof of the effectiveness of awareness methods from psychological theories and they are still silent on the fundamental assumptions of these methods. A wide form of completely different strategies is

now being adopted to measure awareness efforts. However; organisations seem to find it difficult to implement effective quantitative metrics [7]. They tend to adopt different methods, both quantitative and qualitative approaches, to measure the effectiveness of their awareness activities. Nyabando agrees that more extensive qualitative and quantitative studies are needed to understand the disparities between awareness and practice [20].

In 2005, a prototype model [15] was invented by Kruger and Kearney to measure awareness effectiveness in an international gold mining company based on knowledge, attitude and behaviour (KAB). However, their work failed to study the basic theory behind the model. Similarly, Hagen, Albrechtsen and Hovden analysed responses to research questions from 87 Information Security managers in Norwegian organisations [9]. Furthermore, Albrechtsen and Hovdenin [2] identified Information Security related discussion as a tool effective enough to raise users' awareness.

Johnson, in his doctoral research [13] conducted at the University of Lagos, argues that too much is expected from the audience, undermining a fact that security processes can only be effective when audience have a good security support and appreciate security requirements. On this basis and by applying background training, the authors in [12] were able to prove that it is very easy (through SN in particular) to capture huge amount of data for effective phishing attacks. However, they attempted (with no success) to measure the influence of social context information on phishing attacks. What makes their work different was that e-mails were spoofed to deceive users as if it was from friends in the Social Networks (SNs), and at the end the total number of victims to this phishing attack outweighed the expectation [12].

Wolf in [30] reported that there were some unconventional methods used to study and measure users' awareness. One of the notable methods was employed by Dodge, Carver, and Ferguson who used phishing e-mails to detect users that clicked on potentially malicious links in e-mails [6]. Briggs made another attempt by describing how software was implemented to examine network traffic for Personally Identifiable Information (PII), and being transmitted unencrypted over a campus network

[4]. Meister and Biermann, using similar methods, detailed the use of a worm that was created to test the users' ability to detect phishing attempts in their research [17]. Nagy and Pecho tested Facebook users' ability to detect and measure phishing attacks by attempting to friend as many unknown people as possible [19].

The study in [4] highlighted 12 different metrics as effective in measuring the success of awareness activities, all of which are *incident statistics* driven. The most popular overall is the measure of internal protection, where policy breaches from audit report are being used as a measure. This is followed by the effectiveness and efficiency measure, where experience of the respondents on security incidents count a lot. The common metrics include the quantity of incidents resulted from human unsecured deeds which are the root cause of most terrible incidents. Surveys and questionnaires are adjudged to be the most popular measurement instrument as evident by the large number of studies that used them. However, Bulgurcu, Cavusoglu and Benbasat in [5] are the only authors who use author observation as part of the measurement. They combined surveys, interviews, case studies and observations to form their analysis [30].

While considering a *non-incident statistic approach* to measure awareness on SNs, [21] in his doctoral research work, identified several applications and utility software presently available to crackSN's password file, but with a limitation to only the files that are hashed with less complex system such as MD5. These password crackers include but are not limited to Facebook Password Sniffer, John the Ripper, Password Decryptor, Google Password Decryptor, Password Security cracker, Password Fox, Sniffer, OperalPassView, Access Pass View, Web PassView, and AsterWin IE. This limitation calls for the need for a stronger password cracker to crack SN password shashed with a more complex function.

3. INITIAL THEORETICAL FRAMEWORK

Sekaran and Bougie define a theoretical framework as the foundation on which a study is

based. They argue that "the relationship between the literature review and the theoretical framework is that the former provides a solid foundation for developing the latter" [25]. They finally recommend that a theoretical framework should have the following three basic features:

- a) Definition of variables considered relevant to the study.
- b) A conceptual model that describes the relationships between the variables.
- c) A clear explanation why these relationships are expected to exist.

Therefore, this study relates these features to the following sections:

3.1 Relevant variables to this study

ENISA in [8] recommends that an information security programme of any organisations (including SNs) cannot be improved upon if the effectiveness of their implemented awareness techniques is not adequately measured. Since the subject being studied is the measurement of awareness techniques using an approach that is not based on incident statistics, this paper focuses on awareness measurement as a global theme. This global theme is a dependent variable since the measurement of awareness is dependent on so many factors, some of which are independent variables for this study.

The independent variables considered relevant to this study include: privacy risks, security threats, security techniques, measurement techniques/approaches, problems in measurement, what to measure, how to measure, and awareness benchmarking and metrics. This is because while the global theme (awareness measurement) is dependent on independent variables in this study, these independent variables do not depend on any factor.

3.2 The conceptual model relating the variables

The conceptual model that relates both the dependent and independent variables, which influence awareness measurement in this study, is provided in figure 1.

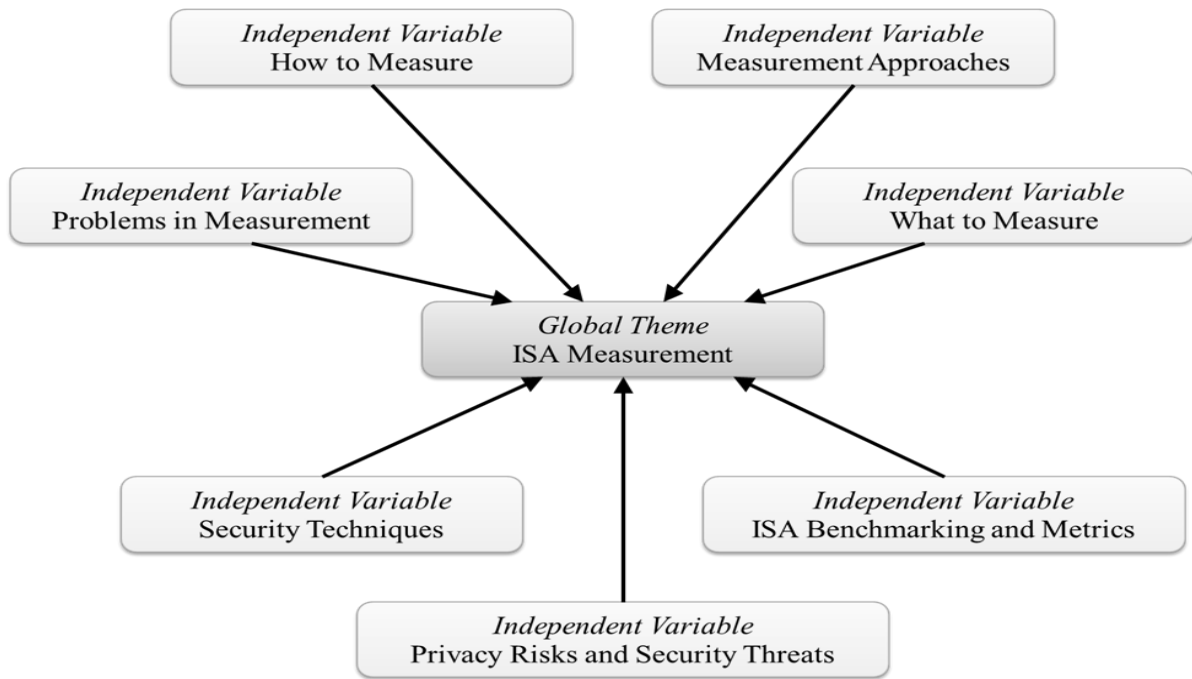


Figure 1: Initial framework for measuring awareness (own compilation)

3.3 Claims for the existence of the expected relationship.

The past related works recognise privacy risks and security threats, comprising cyber-attacks and users’ identity theft, as factors to be curtailed by the privacy approach when raising awareness [21]. Security techniques as well as available measurement approaches are also identified as potential factors influencing awareness measurement, and are therefore considered as independent variables in this study.

Similarly, literature findings in [21] point to the possible influence of what to measure, how to measure, awareness benchmarking and metrics, determinant factors, and problems in measuring the effectiveness of awareness efforts. These factors are the themes or relevant variables highlighted in section 3.1.

4. THE REFERENT THEORIES

The referent theories for this study are the Markus’s three theories as presented by Myers and Avisonin [18], namely: system-determined,

interaction-determined and people-determined. This research on measuring awareness efforts in SNs based on the concepts from dissimilar models applied to the field of information security and human behaviour, namely: activity theory (AT), knowledge, attitude and behaviour (KAB).

Hashim and Jones employed AT to investigate students’ response to information security on their learning system and suggest that AT theory is applicable to human beings and their technological environment [10]. Similarly, the users’ KAB must be in line with the security requirement for the SN to be adequately secured [16]. However, since the required changes in security behaviour on SN may not be easily attained using the KAB model alone [21], the combination of both the AT theory and KAB model is used to guide this study. Figure 2 illustrates the AT model as applicable to this study.

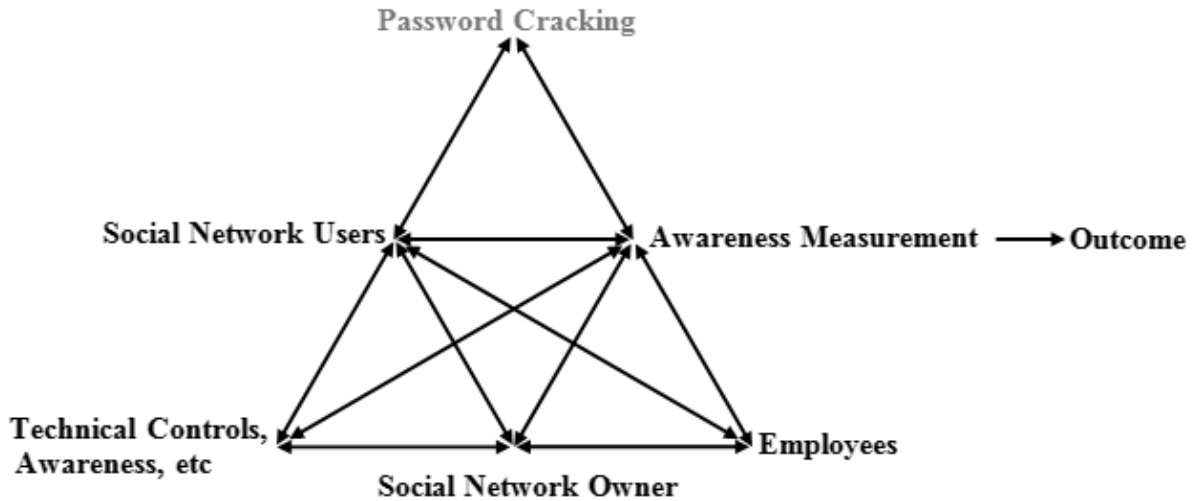


Figure 2: Activity theory for this study

Relating the AT model to this study, the subjects are ‘SN users’ who are affected by the technical security on the network and are also expected to radiate secured behaviour and attitude. The object is the ‘awareness measurement’, which is the activity under study, and the tool or instrument is the ‘password cracking’, which is the approach that is not based on incident statistics, developed to proactively measure the awareness efforts in SNs. The community is the ‘SN’s owner’ who implemented rules (‘technical controls’, ‘privacy policy’, ‘awareness’, etc.), while the division of labour refers to the ‘employees’ of the SN who are saddled with segregation of duties foreffective internal security on the SN.

5. METHODOLOGY

In this study, quantitative research strategy is deployed where data-gathering (Figure 3) primarily focuses on collection of existing data already gathered by sOcialistOnline, a Social Network newly developed by Okesola in [21]. These existing data are relevant data from management information systems such as users’ personal and confidential data including photographs and password files.

To ensure data collected from the study is treated as private and confidential, and to comply with the ethics policy on research, the crackeris designed to display *only* the security information

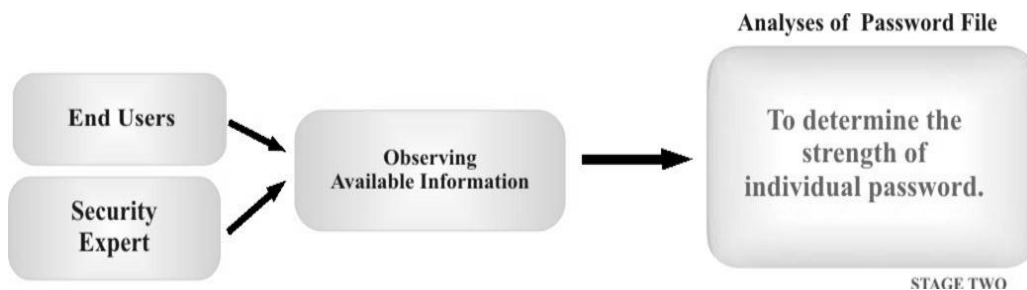


Figure 3: Data Gathering

about passwords without actually showing the password itself (see figure 4).

5.1 Developing the cracker

This newly developed cracker, which is anutility software, can also crack and display security information of the passwords (and not the

password itself) stored on Microsoft Outlook, Mozilla Firefox, and Internet Explorer and displays password security information such as password strength. This information shall be used to determine the strength of passwords used by the users of sOcialistOnline, without necessarily seeing the passwords themselves.

Some of the features of the cracker are described as follows.

- 1. System requirements/security:** This cracker is restricted to work only on Window 2000 version and up to Window 7, which are operating systems whose security settings are considered effective. Although hashed with MD5 system which is currently regarded as indecipherable, the newly developed cracker is stored and run offline from an external disk to wall it off from possible exploitations by a desperate hacker.
- 2. Applications supported:** This cracker was incidentally tested and found suitable to crack the passwords of Internet Explorer 7.0 - 9.0, Internet Explorer 4.0 - 6.0, and Microsoft Outlook as well.
- 3. Known limitations:** Only two limitations were noticed: (1) once protected by a master password, this cracker cannot crack Firefox passwords; and (2) Windows passwords can only be uncovered if the cracker is run with administrator's privileges.
- 4. Columns description:** The cracker output has eight columns namely: user name, uppercase, lowercase, numeric, special, password length, repeating, and password strength. The columns are displayed in figure 4 and described as follows:

User Name	Uppercase	Lowercase	Numeric	Special	Password Length	Repeating	Password Strength
Johnson	0	0	5	0	5	0	Very Weak
admin	0	0	5	0	5	0	Very Weak
olayemi	0	6	0	0	6	0	Very Weak
jezz	0	6	0	0	6	0	Very Weak
ojoja	0	7	1	0	8	0	Medium
jamopow	1	7	1	0	9	0	Strong
owoblow	0	9	0	0	9	0	Weak
messi	0	0	10	0	10	0	Weak
rolex	1	7	1	0	9	0	Strong
adex	1	7	1	0	9	0	Strong
neyor	2	10	1	2	13	2	Very Strong

Figure 4: The screen-print of the password cracker

User name: The user name or UserID of the particular password item.

Uppercase: The total number of characters with uppercase (A - Z) in a password.

Lowercase: The total number of characters with lowercase (a - z) in a password.

Numeric: The total number of numerals (0 - 9) in a password.

Special: The total number of characters that are non-alphanumeric in a password.

Password Length: The total number of letters or characters in a password.

Repeating: The total number of characters repeated in the password. For instance, if the password is cnbncck, the repeating value will be two since only c and n characters appear more than once.

Password strength: This may be calculated based on the total number of parameters including the character type, the presence and the total number of characters and repeating characters used in the passwords.

Each value appearing in this column denotes the strength of the password as classified in table 1.

5.2 Password classification

For this research exercise, a password is

classified as either good or bad. Going by table 1 therefore, a password combination is said to be:

- *Bad* and *guessable* if its character combination is *weak* or *very weak*; and
- *Good* if its character combination is *medium*, *strong*, or *very strong*.

Table 1: Password Classification

Security classes	Sub-classes	Class Interval	Character composition
BAD	Very weak	1 – 5	Less than 8 characters length, only alphabets or numbers.
	Weak	6 – 15	Only alphabets or numbers but longer than 7 characters.
GOOD	Medium	16 – 29	Alphabets (lowercase or uppercase) plus numbers and longer than 7 characters.
	Strong	30 – 49	Uppercase, lowercase plus numbers and longer than 7 characters.
	Very strong	50 – 9999	Uppercase, lowercase, numbers, plus special characters (#, \$) and longer than 7 characters.

Following this classification, data related to the demographic details were captured and analysed to generate a survey report. These data include Group, age, gender, tribe, country, qualifications, profession, and technological advancement based on the user level of computer literacy and proficiency.

5.3 Performing the cracking

The encrypted sOcialist Online password file was decrypted and downloaded into a flat file. Personal data of all the users on the SN were captured but bearing in mind that most people who join will not remain active for privacy, social, and other factors [21], the file was

automatically reviewed to eliminate the non-active users. The cracker statistically analyses the compositions of the *active* passwords to determine their strengths and weaknesses as *very weak*, *weak*, *medium*, *strong*, or *very strong*.

6. FINDINGS AND DISCUSSION

The result from the password cracking is summarised in table 2, where the numbers of good and bad passwords are almost at ratio 5:1 with bad password taking less than 20% of the total population size. As indicated in table 2 and figure 6, only 364 passwords (representing 19%) of the population size of 1,903 are bad passwords.

Table 2: Output from the Password Scanner

Password Classes	Good	Bad	Population Size
Very Weak		135	135
Weak		229	229
Medium	384		384
Strong	397		397
Very strong	858		758
TOTAL	1,639	364	1,903

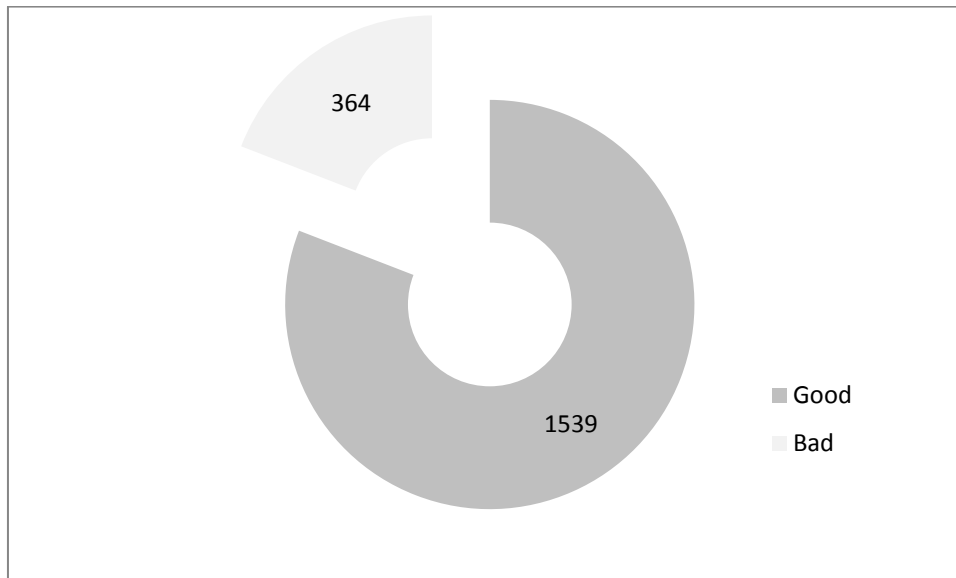


Figure 6: Good vs. bad password combinations (own compilation)

This is an improved result on Samuel and Samson findings in [24], where the relationship between good and bad passwords was almost at par. The success recorded in this study (where percentage of good password combination is far higher than bad password combination) could be traced directly to various adequate and effective awareness techniques specifically implemented on sOcialistOnline. These include Explicit Privacy Policy, Privacy Awareness and Customisation, Data Minimisation, Privacy Lens, Password Monitoring and Standardisation [21].

6.1. A proactive measurement approach

The major achievement here, which remains the main focus of this research, is that the effectiveness of awareness efforts implemented on the SN was successfully measured using a technique that is not incident statistics driven. Hence, the guessable password combinations could easily be discovered before the occurrence of an event or success of an attack.

This is an improvement on some past related works of Briggs [4], Carnegie Mellon University [26], and CISO - a large finance service German organisation [29], where awareness simprints were evaluated on the basis of post-mortem metrics. Although the statistics generated from these incident-statistics approaches are always of great interest to senior management, the techniques may still not be most effective because it cannot give a true reflection of awareness[4].

6.2. Theoretical contributions

The theoretical contribution of this research is the adaption of the underlying and replicating theories to study the security awareness measurement on SN as follows:

6.2.1 Adapting assumptions of underlying theories

This author specified the main referent theories on which the study is based and identifies other unrelated models but practically similar to this study. The assumptions underlying systems security and system utilization are *adapted* from these referent theories of system and people's resistance to study the impacts of awareness on SN. The facts in the 'real world' are the themes of this study, which are those variables influencing the measurement of awareness efforts in SNs and are grouped into seven categories for the purpose of this study

6.2.2 Replicating theories from other domains

Theories from other domains, including human behaviour and psychology are replicated in this study and applied to the field of awareness measurement. This work replicates the general concepts adopted in models such as AT, KAB and system-determined and people-determined theories, which are technologically applicable and widely accepted.

7. CONCLUSION AND RECOMMENDATION

The goal of password cracking in this study is to provide a useful direct quantitative measure of the attitude and behaviour of SN users. Following the successful cracking of users' password files, the proposed solution in this study analyses the strength of individual passwords, using an automated statistical approach. The number of users using easily guessable passwords is a key indicator of effective awareness [8].

People generally do not see legitimate reasons behind the creation of password cracker.

However, the problem is actually not the existence of password crackers, but their frequent illegal use by fraudulent people for bad goals and objectives. When employed for good intentions, password crackers can offer a valuable service to system and data administrators by alerting them of system or users' weak passwords [27]. The stronger password cracker developed in this research is thereby recommended for use to measure the effectiveness of awareness efforts on the SNs, as it is capable of cracking SN passwords that is hashed with a more complex system than MD5.

REFERE3NCES

- [1] ABC (2013). Measuring Information Security Awareness Techniques; the Pros and the Coins. Academy. Press, Port Harcourt, Nigeria
- [2] Albrechsten E. and Hovden J. (2010). Improving information security awareness and behaviour through dialogue, participation, and collective reflection. An Intervention Study. *Computer & Security*, 29(4): 432-445
- [3] Brodie, C. (2009). The Importance of Security Awareness Training. SANS Infosec Reading Room. Available online: http://www.sans.org/reading_room/whitepapers/awareness/importance-security-awareness-training_33013. Date Retrieved: May 23, 2011.
- [4] Briggs, L. (2009). Locking Down Data at U Nebraska. Available online: <http://campustechnology.com/articles/2009/10/15/locking-down-data-at-unebraska.aspx>. Date Retrieved: July 23, 2011
- [5] Bulgurcu, B., Cavusoglu, H. and Benbasat, I. (2009). Effects of Individual and Organisation Based Beliefs and the Moderating Role of Work Experience on Insiders' Good Security Behaviours. 2009 International Conference on Computational Science and Engineering
- [6] Dodge, R. C., Carver, C. and Ferguson, A. J. (2007). Phishing for User Security Awareness. *Computers and Security*, 26, 73-80
- [7] ENISA (2007). Information Security Awareness Initiatives: Current Practice and Measurement of Success. Available Online: http://www.itu.int/osg/csd/cybersecurity/WSIS/3rd_meeting_docs/contributions/enisa_measuring_awareness_final.pdf. Date Retrieved: May 18, 2011.
- [8] ENISA(2008). Information Security Awareness Initiatives: Current Practice and Measurement of Success. Available Online: http://www.itu.int/osg/csd/cybersecurity/WSIS/3rd_meeting_docs/contributions/enisa_measuring_awareness_final.pdf. Date Retrieved: May 18, 2011
- [9] Hagen, JM, Albrechtsen E., and Hovden J. (2008). Implementation and Effectiveness of Organisational Information Security Measures. *Information Manage. Computer Security*, 16(4):377-397
- [10] Hashim, N.H. and Jones, M.L. (2007). Activity Theory: A framework for qualitative analysis. Research online, 4th international Quality Research Convention (QRC), Hilton, Malaysia. Available online: <http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1434&context=commpapers>. Date Retrieved: September 14, 2014.
- [11] Hinson, G. (2012). The True Value of Information Security Awareness. Noticebored. Available Online: http://www.noticebored.com/html/why_awareness_.html. Date Retrieved: August 23, 2011.
- [12] Jagatic, TN., Johnson, M., Jakobsson, M., Menczer, F. (2007). Social Phishing.

- Communications of the ACM. Vol. 50(10) 112-136.
- [13] Johnson, A. (2012). Social Network Settings are Ineffective. *Information and Communication Journal*, Department of Computer Sciences, University of Lagos, Nigeria. Vol. 12(3) 25-34
- [14] Khan, B, Alghathbar, K.S, Nabi, S.I, and Khan, M.K. (2011). Effectiveness of Information Security Awareness Methods Based on Psychological Theories. *African Journal of Business Management* Vol. 5(26). Available online: http://www.academicjournals.org/ajbm/pdf/pdf2011/28_Oct/Khan%20et%20al.pdf. Date retrieved: October 27, 2012.
- [15] Kruger, H. and Kearney (2005). Measuring Information Security Awareness: A West Africa Gold Mining Environment Case Study. A Proceedings Of ISSA, Pretoria, South Africa. Available online: http://icsa.cs.up.ac.za/issa/2005/Proceedings/Full/018_Article.pdf. Date Retrieved: July 25, 2011.
- [16] Lacey, D. (2009). Managing the Human Factor in Information Security: How to win Over Staff and Influence Business Managers. Wiley, John and Sons, Incorporated: Hoboken, New Jersey, ISBN: 0470721995, ISBN-13: 9780470721995, 384 pages.
- [17] Meister, E. and Biermann, E. (2008). Implementation of a Socially Engineered Worm to Increase Information Security Awareness. Third International Conference on Broadband Communications, Information Technology and Biomedical Applications, Gauteng.
- [18] Myers, M.D., and Avison, D. (2002). Qualitative Research in Information Systems: a reader. (Michael D. Myers and David Avison., Ed.). London: Sage Publications, pp 1-312.
- [19] Nagy, J. and Pecho, P. (2009). Social Networks Security. The Third International Conference on Emerging Security Information, Systems and Technologies, Greece. Date Retrieved: August 3, 2012
- [20] Nyabando, C. J., (2008). An Analysis of Perceived Faculty and Staff Computing Behaviours That Protector Expose Them or Others to Information Security Attacks. Doctoral Dissertation, East Tennessee State University. Available Online: <http://proquest.umi.com/pqdlink?did=1597602021&fmt=2&vtype=PQD&vinst=PROD&RQT=309&vname=PQD&TS=1316378955&clientid=79356>. Date Retrieved: September 18, 2011
- [21] Okesola, J.O. (2014). Measuring Information Security Awareness Effectiveness in Social Networking Sites – A Non-Incident Statistics Approach. PhD Thesis (UNPUBLISH\ED), School of Computing, University of South Africa (UNISA).
- [22] OnlineHashCrack.com (2013). The Truth About Facebook Password Hacking. Available online: http://www.onlinehashcrack.com/how_to_crack_facebook_account_the_truth.php. Date Retrieved: March 5, 2013
- [23] Price Water House Coopers (2010). Protecting Your Business – Security Awareness: Turning Your People into Your First Line of Defence. Available online: http://www.pwc.co.uk/eng/publications/protecting_your_business_security_awareness.html. Date Retrieved: July 25, 2011.
- [24] Samuel, J. and Samson, P. (2012). Unsecured Users' behaviour on the SNS; the outcome of Poor Password Combinations. Symposium on Usable Privacy and Security (SOUPS'12), Pittsburgh, PA, U.S.A.
- [25] Sekaran, U. and Bougie, R. (2010). Research Methods for Business: A Skill Building Approach (5th ed.). Chichester: John Wiley & Sons, pp 1-488.
- [26] Spice, B. (2007). Carnegie Mellon Researchers Fight Phishing Attacks With Phishing Tactics. Available Online: http://www.cmu.edu/news/archive/2007/october/oct2_phishing.shtml. Date Retrieved: December 21, 2012
- [27] Taber, M. (2011). Maximum Security: A Hacker's Guide to protecting Your Internet Site and Network. Macmillan Computer Publishing. Available online:

- <http://newdata.box.sk/bx/hacker/ch10/ch10.htm>. Date Retrieved: April 16, 2013
- [28] Veseli, I. (2011). Measuring the Effectiveness of Information Security Awareness Programme. Master Thesis, Master of Science in Information Security, Department of Computer Science and Media Tech., Gjovic University College. Available online: http://brage.bibsys.no/hig/bitstream/URN:NBN:no-bibsys_brage_21083/1/Iilirjana%20veseli.pdf. Date Retrieved: June 5, 2012
- [29] Williams, A. (2007). The Ineffectiveness of User Awareness Training. Available Online: <http://techbuddha.wordpress.com/2007/05/02/the-ineffectiveness-of-user-awarenesstraining/>. Date Retrieved: March 3, 2011
- [30] Wolf M.J. (2010). Measuring Information Security Awareness Programme. Master's Thesis, Department of Information Systems and Quantitative Analysis, University Of Nebraska, Omaha.

IMPACTS OF SECURITY DIMENSIONS ON AWARENESS MEASUREMENT IN SOCIAL NETWORKING

J. O. K. Okesola¹, A. A. Owoade², O. W. Adesanya³, F. M. Babalola⁴

¹Covenant University, Ota, Nigeria

²Tai Solarin University of Education, Ijebu-Ode, Nigeria.

³Federal College of Agriculture, Akure, Nigeria.

⁴The Polytechnic, Ibadan, Nigeria.

¹48948535@mylife.unisa.ac.za; ²owoadeakeem@yahoo.com;

³Sanyalanre2003@gmail.com; ⁴floxymbabs@ymail.com

ABSTRACT

A big challenge facing Social Networks (SNs) and other organisations has been what to measure when determining the adequacy and effectiveness of awareness programmes. This study defines security dimension as Knowledge, Attitude and Behaviour, and identifies them as the main influencing factors to consider in awareness measurement. Web quiz was developed for data gathering and risk scores were calculated to validate the research findings by investigating the impacts of these factors on awareness in SNs.

Keywords: Information security awareness, non-incident statistics, password cracker, risk score, social networks.

1.0 INTRODUCTION

Information Security Awareness (awareness) programme may be successfully implemented and even carry top management supports, yet organisations cannot be too sure that every stakeholder understands his security roles and responsibilities [17]. While it is difficult and may be deceitful to assume the success of any security effort, it is pleasing and more assuring to measure the status of awareness programmes on the Social Network Sites (SNs). Hence, a more structured approach is required to study the effects of awareness techniques on the SNs in order to ascertain its contribution to the field of security [7].

Authors and managers are often confronted with two distinctive challenges when it comes to developing a measuring tool and subsequently, performing the measurement.

These challenges have to do with *what to measure* and *how to measure* [7], and have been found to be promoted by certain requirements such as sustainability, the use of scientific methods, ease of use, and compliance to organisation's requirements. Okesola in [14] used the combination of quantitative and qualitative methods to measure awareness efforts using a non-incident statistics approach. He concludes that technical controls may be used to measure the effectiveness of awareness efforts in SNs only that the findings must be validated to eliminate the possibility of *influencing factors* on the final results.

According to [14], the influencing factors are regarded as the three security dimensions and include *knowledge* (that focuses on what a user knows); *attitude* (which focuses on a user's mind-set); and *behaviour* (which

focuses on what a user does). Hence most awareness solutions are built around the Knowledge, Attitude and Behaviour (KAB) model which are principally centred on human knowledge. The model postulates that knowledge accumulation in a particular behaviour (information security, health, environment, education, etc.) instigates a change in attitude. It principally substantiates the influence of knowledge over people's behaviour and knowledge acquisition, which in turn enforces changes in attitude and behaviour. Therefore, for SNs managers to be sure their awareness effort is effective, the impacts on these three elements on its research findings have to be adequately measured [14].

What to measure is crucial but in practice, it is somehow tricky to identify the right metrics [5]. Security measurement is all about common sense; managers must know what to measure, arrange them in a meaningful and manageable order, and come up with a repeatable formula to disclose the security status and how this status changes over time [9].

Generally speaking, when classifying what to measure, the three dimensions of KAB model – *knowledge*, *attitude*, and *behaviour* should be absolutely considered [18]. Figure 1 presents a funnel of these KAB dimensions. As applied in the Venn diagram, SN will be better secured when the KAB of users are tailored towards the security requirements and objectives [8].

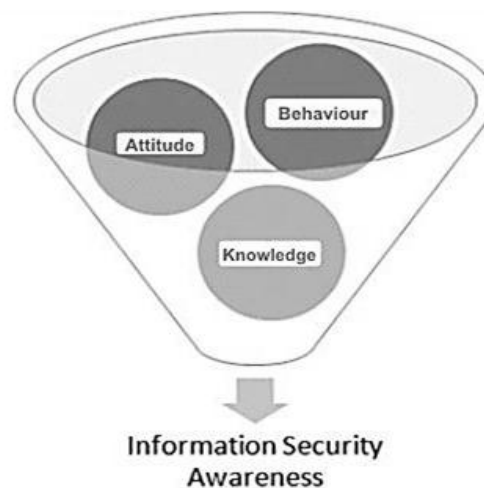


Figure 1: Dimensions of awareness

Knowledge – It is important because a user cannot carry that intention into action without the adequate knowledge and understanding, even if he/she believes security is important.

Attitudes - Unless users have a strong believe in security, they are not going to work securely regardless of their knowledge and understanding of security requirements [8]. Attitude is therefore a dimension that signifies an employee's disposition to act.

Behaviour – Regardless of his knowledge, an individual is not going to impact security unless he/she exhibits some secured behaviours [14]. Coming from the users' end therefore, security awareness is a combined function of knowledge, attitudes, and behaviours. Awareness solutions are typically built around KAB model which centred on human knowledge [7]. That is, the knowledge accumulation in a particular behaviour (information security, health, environment, education, etc.) instigates a change in attitude. The model principally substantiates the influence of knowledge over peoples' behaviour and knowledge acquisition, which in turn enforces changes in attitude and behaviour [6].

2.0 RELATED WORK

Some researchers have been focusing on what users believe and do about security in the real world. At the same time other researchers have been looking at security in the organisational context, keeping in mind requirements of the organisation and user participation to support compliance [18]. Kruger and Kearney in [7] give an example on developing a model for measuring information security awareness. This model was applied in an international gold mining company, with the goal to monitor changes in security behaviour. As a result security awareness campaigns are reviewed whenever there is a need.

Other related studies such as [10] and [12] have also confirmed that knowledge can shed light onto a change process if incorporated into a conceptual framework. However, a change in user's behaviour is not necessarily limited to a rise in knowledge, as behavioural change is a function of multiple variables [6]. Hence the required attitudinal and behavioural changes may not be easily attained or sustained by the use of KAB model alone. A prototype model was invented by Kruger and Kearney to measure awareness effectiveness in an international gold mining company based on knowledge, attitude and behaviour [7]. However, their work failed to study the basic theory behind the model. Furthermore, Albrechsten and Hovden identified an information security related discussion as a tool effective enough to raise users' awareness, although their study approach is comparatively adjudged to be ineffective [1].

Similarly, Wolf in [19] reported that there were some unconventional methods used to study and measure awareness in SN's users. One of the notable methods [3] was employed by Dodge, Carver and Ferguson, who used phishing e-mails to detect users that clicked on potentially malicious links in e-mails. Briggs in [2] made another attempts by describing how software was implemented to examine network traffic for Personally Identifiable Information (PII) that

was being transmitted unencrypted over a campus network. Using similar methods, Meister and Biermann detailed the use of a worm that was created to test the users' ability to detect phishing attempts in their research [11]. Nagy and Pecho [13] tested Facebook users' ability to detect and measure phishing attacks by attempting to befriend as many unknown people as possible. However, all these methods are adjudged to be incident statistics, since the measurement of their effectiveness is based on the occurrence of an event or success of an attack.

Okesola in [14] eventually introduced a proactive approach of measuring awareness efforts in SNs. He designed a secured SN-sOcialistOnline - and implemented a password cracker to determine the impact of various awareness techniques on SN before the occurrence of an event or success of an attack. The study established that awareness efforts have absolute control on user knowledge but not necessarily on the control metrics (habit, attitude, intention, and behaviour). Therefore, since control metrics are deterrents to awareness efforts, [14] admitted that future work is required in this area to put the metrics under controls.

3.0 METHODOLOGY

An in-depth literature study could not produce any questionnaire suitable to capture awareness contents of SN's users that include normative beliefs and habit. Surveys such as [4], [10], and [16] have been done on awareness but these questionnaires did not offer the information and data needed to explore relationship between users' Knowledge, Attitude, Behaviour (KAB) and awareness. Surveys from security awareness metrics were found useful to some extent; in particular that of Mitchamzin [12]. However, none of these research works covers the scope of this study and only individual questions could be recycled.

3.1 Developing the Web Quiz

The author consulted related literature and regarding human behaviour towards the use of SNs to compile a questionnaire, which was transposed to a web quiz for this study. Data related to the measurement dimension of awareness and some other following facets were captured and analysed:

- **Knowledge:** Password handling, virus prevention and controls based on the users' basic understanding of security perspectives as related to awareness.
- **Attitude:** SNs' settings and self-hiding of profile data are important security issues on SNs, which a user may or may not believe in. It is purely based on the user's belief about information security safety and privacy on the SN and the priority a user gives to security type (economic, reputation or physical).
- **Behaviour:** How the user protects his sign-on, safeguards information, controls and sets his privacy, reads and understands the SN's policy before signing on. It may also encompass users' reaction to general e-mail and internet issues with respect to awareness. These are user styles of living or work which is insecure upon his adequate knowledge of awareness

3.2 Conducting the Survey

The quiz template was administered to participants with guessable password combinations, requesting for their responses. Given that the survey can be conducted with the administration of either the *questionnaire* or *interview*, the author chooses to apply the four principles of contextual interviewing to

guide the quiz administration process. These principles include context, interpretation, partnership, and focus [15]. In this study, all the participants were exposed to the same set of questions. Through a covering memo popping out at the first page of the quiz template, all participants were informed that the exercise is optional and that its objective was to ascertain that their insecure behaviour on the SN is not attributed to inadequate awareness efforts. The information gathered at this phase was analysed further to form an opinion.

3.3 Measuring KAB

The survey is designed to measure KAB, which are the three dimensions of awareness. This section surveys user's privacy priority, password security management, confidence in existing SN's settings, compliance to SN's policy, users' habits and behaviour. Questions 1 to 15 representing 75% of the questionnaire address KAB while only 25% is on intention and some other factors. The survey tests the knowledge, attitude and behaviour of SN's users towards security related questions and situations.

Optional answers to some of these multiple questions denote strong awareness and good practices while others signify insecure attitudes and behaviour of higher-risk activities. On this basis therefore, *risk/significant* value is assigned to each question's response, ranking from 1 to 4 with "one" being the lowest and "four" being the highest risk/significant level. When collected and properly analysed, the result of the *significant level* was used to determine the influence of each of the KAB elements on awareness following the following steps:

1. Multiply respond risk value (1-4) of each of the 20 questions by the number of participants that answer it. That is,
 $\{Response\ risk\ value\} \times \{no\ of\ times\ chosen\} \Rightarrow Response\ Total \dots\dots\dots equation\ 1$
2. Sum-up the Response Total (RT) for each of the questions to obtain the Cumulative Response Total (CRT) value.

$$\sum_{k=1}^5 (RT)k = CRT \dots\dots\dots equation\ 2$$

- Use the number of time chosen (NTC) to divide the survey *cumulative* response total (CRT) to obtain the SN’s Risk Score (RS) for each question.

$$\frac{CRT}{NTC} = RS \dots \dots \dots \text{equation 3}$$

(Risk score is the probability of a user compromising the system, and it is used in this study to determine the validity of the research findings).

- Sum the risk score per subject area to obtain the *cumulative* risk score (CRS) for each of the KAB dimension and intension.

$$\sum_{j=1}^4 (RS)_j = CRS \dots \dots \dots \text{equation 4}$$

- Using the calculated *cumulative* risk score (CRS), cross-check the *Risk Levels* in Table 1 to determine the SN’s general risk rating.

Table 1: The Potential Risk Level

Risk Levels	Range	Description
Low	1 – 5	SNs’ users are fully aware of security policies and procedures, the potential risks and possible consequences of a threat. They have all the training and exposure required to mitigate an attack.
Elevated	6 – 8	Users have a good understanding of awareness with adequate exposure but they may choose not to comply with good security policies, principles and controls.
Moderate	9 – 12	Users are aware of security threats and potential attacks but they have a knowledge gap in identifying or reporting security events and therefore require training and exposure.
Significant	13 – 16	Users understand security policies and standards but they do not believe they are truly accountable for their security performance. They assume security lies on technology and the management of the SN.
High	17 – 20	Users are not interested in controls and security for reasons only known to them. They exhibit risky behaviour that can be easily exploited, thereby making the SN more vulnerable to various attacks.

Since each *risk level* contains five questions with a minimum score of one and maximum score of four, then

Minimum Risk Score = (5 x 1) = **5**; and Maximum Risk Score = (5 x 4)=**20**.

4.0 RESULTS AND EVALUATION

4.1 The sample size

Out of the total population of 2,436 users on sOcialistOnline as at July 8, 2013, only 2,015 thousand users were *active* representing ratio 1:5 against *inactive* participants. In other to obtain a more reliable result and to abide by government regulations regarding children completing questionnaire, the survey was restricted to only the teenagers and matured participants;

in which case, 112 users below the age of 13 years were further eliminated from the active sample size. The study therefore has a final population size of 1,903 active participants (adults and teenagers), 66% male and 34% female. Mostly students from Nigeria and few other African countries, the participant ages lie between 13 years and 53 years, with an average age of 22 years. These population distributions of the SN’s users are as represented by the pie charts in figure 2.

a.) Active Vs. Inactive.

b.) Adult Vs. Infants.

c.) Male Vs. female.

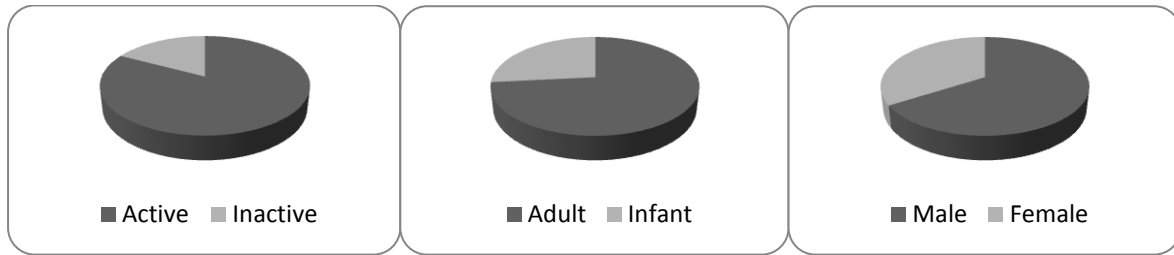


Figure 2: SNS's Users Population Distribution Ratios

5.0 THE FINDINGS

Table 2 depicts the results from the survey conducted on SN users with guessable passwords based on the questionnaire/quiz presented to them. The population size of this class of bad passwords is 364. However, since this exercise was voluntary and no incentive was provided to the participants, 52 participants declined to respond to the quiz thereby bringing the class size to 312. This quiz consists of only 20 questions, 5 each from knowledge, attitude, behaviour, and users' needs and objectives. The column description for the result is as presented on the table below.

Table 2: Results from the questionnaire

	<i>Question's Nos.</i>	<i>No of time Chosen (312)</i>	<i>Cumulative Response Total</i>	<i>Risk Score</i>
Knowledge	1	312	404	1.29
	2	295	313	1.06
	3	302	442	1.46
	4	245	295	1.20
	5	278	287	1.03
	CRS (K)			
Attitude	6	285	1,049	3.68
	7	312	897	2.88
	8	291	952	3.27
	9	278	992	3.57
	10	287	778	2.71
	CRS (A)			
Behaviour	11	248	384	1.55
	12	312	1,120	3.59
	13	289	853	2.95
	14	307	850	2.77
	15	311	1,145	3.68
	CRS (B)			

- First column: the description of what a particular question is all about.
- Second column: the serial number of the quiz questions specifically meant to identify the question.

- Third column: the total number of time each of the questions was answered by the respondents. It is surprising to note that only 2 questions 1 and 12 were fully attended to by all the **312** participants; the rest are far below 312 with question 4 being the least patronage
- Fourth column: this is the column that warehouses the summation of response total for each of the questions. The computation of both Response Total (RT) and Cumulative Response Total (CRT) are done with equation 1 and 2 respectively.
- Fifth column: this is the column for total Risk Score per survey question. The RS is computed using equation 3 and the summation of the total Risk Score (RS) per survey question gives the total calculated cumulated Risk Score (CRS) for each security dimension. This is denoted as CRS(K), CRS(A), CRS(B) and CRS(I) for Knowledge, Attitude, Behaviour and Intention respectively, and are compared with *Risk Levels* on table 1 to measure the general risk rating of the SN.

Relating the computed CRS in Table 2 to the ranges of potential risk levels on the SN, it is interesting to note that *knowledge* is the only security dimension that does not pose a serious security issue. Knowledge with $CRS(K) = 6.04$ presents *low/elevated risks* to the SN, implying that users generally have a good understanding of awareness with adequate exposure, which is as a result of various awareness techniques implemented on the SN – sOcialilstOnline. Therefore, since $CRS(K) < 9$ (Table 1 & Table 2), users' *knowledgewhich isa security dimension*, is directly influenced by awareness impact in SN.

Similarly on Table 2, $CRS(K) > 12$ and $CRS(B) > 12$, each of which is high and significant enough to pose a serious security threat on any Social Network. Despite the adequate awareness techniques implemented on sOcialistOnline, which boosted users'

knowledge as evidenced by $CRS(K) < 9$, users' insecure *attitude* and *behaviour* make the efforts ineffective. Some users in this category still do not believe that security lies on the users but on technology and the management of the SNs. Other users have faith in security and controls but they are just not interested for reasons yet to be understood. They exhibit risky behaviour that can be easily exploited thereby making the SN more vulnerable to various attacks.

This situation can be likened to a case of a medical doctor that still smokes cigarettes despite his better understanding and beliefs in safe practice and healthy diet. Obviously, a medical doctor must have been well trained and exposed, thereby fully aware of medical implications of smoking before being conferred a medical degree. A lack of adequate awareness (knowledge) is therefore not an issue in this case but habit, beliefs, and style of living.

6.0 CONCLUSION AND RECOMMENDATION

Results obtained have proven that awareness efforts have direct influence only on Knowledge and not necessarily on attitude and behaviour. It then follows that, knowledge can measure only the adequacy (and not the effectiveness) of awareness efforts. Training for instance, can influence only the users' knowledge towards awareness and no more because other factors (attitude and behaviour) are habitual, and may remain perpetual regardless of the training efforts [20]. Hence, users with unpleasant attitude, insecure behaviour and evil intention will disregard their security knowledge and expose the SNs to various threats and vulnerabilities. Therefore, whether or not the implemented awareness effort is strong and adequate, users' attitude and behaviour may still render them useless. Hence, knowledge can only measure the adequacy (and **not** the impacts) of the awareness since awareness can only influence users' knowledge.

REFERENCES

- [1] Albrechsten E. and Hovden J. (2010). Improving Information Security Awareness and Behavior Through Dialogue, Participation, and Collective Reflection. An intervention study. *Journal of Computer Security*, 29(4): 432-445
- [2] Briggs, L. (2009). Locking Down Data at U Nebraska. Available online: <http://campustechnology.com/articles/2009/10/15/locking-down-data-at-unebraska.aspx>. Date Retrieved: July 23, 2011
- [3] Dodge, R. C., Carver, C. and Ferguson, A. J. (2007). Phishing for User Security Awareness. *Journal of Computers and Security*, 26, 73-80.
- [4] Guenther, M (2001). Melissa Guenther, LLC. Available online: <http://www.iwar.org.uk/comsec/resources/sa-tools/Security-Awareness-Quiz-Questions.pdf>. Date Retrieved: February 27, 2013.
- [5] Hinson, G. (2006). Seven Myths About Information Security Metrics. Noticebored. Available Online: http://www.noticebored.com/IsecT_paper_on_7_myths_of_infosec_metrics.pdf. Date Retrieved: February 9, 2011.
- [6] Khan, B, Alghathbar, K.S, Nabi, S.I, and Khan, M.K. (2011). Effectiveness of Information Security Awareness Methods Based on Psychological Theories. *African Journal of Business Management* Vol. 5(26). Available online: <http://www.academicjournals.org/ajbm/pdf/pdf2011/28Oct/Khan%20et%20al.pdf>. Date retrieved: October 27, 2012.
- [7] Kruger, H. and Kearney (2005). Measuring Information Security Awareness: A West Africa Gold Mining Environment Case Study. A Proceeding of ISSA, Pretoria, South Africa. Available online: http://icsa.cs.up.ac.za/issa/2005/Proceedings/Full/018_Article.pdf. Date Retrieved: July 25, 2011.
- [8] Lacey, D. (2009). *Managing the Human Factor in Information Security: How to win Over Staff and Influence Business Managers*. Wiley, John and Sons, Incorporated: Hoboken, New Jersey, ISBN: 0470721995, ISBN-13: 9780470721995).
- [9] Lindstrom, P (2012). Security: Measuring up. *Information Security Magazine*. Available online: <http://searchsecurity.techtarget.com/tip/security-measuring-up>. Date Retrieved: June 21 2011.
- [10] Manly, C (2013). IT Security Awareness Survey. Library System, Cornell University Library Information Technologies. Available online: https://cornell.qualtrics.com/SE/?SID=SV_1Xs2GRV6rWl88Xr. Date Retrieved: May 27, 2013.
- [11] Meister, E. and Biermann, E. (2008). Implementation of a Socially Engineered Worm to Increase Information Security Awareness. Third International Conference on Broadband Communications, Information Technology and Biomedical Applications, Gauteng.
- [12] Mitchamz, Z.S. (2013). Security Awareness Metrics – Informal Survey by HEISC. EDUCAUSE. Available online: <http://preview.tinyurl.com/awarenessmetrics>. Date Retrieved: May 27, 2013.
- [13] Nagy, J. and Pecho, P. (2009). Social Networks Security. The Third International Conference on Emerging Security Information, Systems and Technologies, Greece.
- [14] Okesola, J.O (2014). Measuring Information Security Awareness Effectiveness in Social Networking Sites – A Non-Incident Statistics Approach. PhD Thesis, School of

- Computing, University of South Africa (UNISA).
- [15] Olivier, M. S. (2004). Information Technology Research. A Practical Guide for Computer Science and Informatics (2nded.). Van Schaik Production.
- [16] SANS (2012). Security Awareness Survey. SANS, Securing the Human. Available online: <http://www.securingthehuman.org/media/resources/business-justification/security-awareness-survey.pdf>. Date Retrieved: May 27, 2013.
- [17] Singh, M and Patterh M.S. (2007). Security Functional Components for Building a Secure Network Computing Environment. Information System Security, Vol. 16(6). Available online: <http://0-search.proquest.com.oasis.unisa.ac.za/docview/229583008/13BC99AE1555BAB50DE/1?accountid=14648>. Date Retrieved: January 20, 2013.
- [18] Veseli, I. (2011). Measuring the Effectiveness of IS Awareness Programme. Master Thesis, Department of Computer Science and Media Technology, Gjovic University College. Available online: http://brage.bibsys.no/hig/bitstream/URN:NBN:No-bibsys_brage_21083/1/Iirjana%20veseli.pdf. Date Retrieved: March 5, 2012.
- [19] Wolf, M.J. (2010). Measuring Information Security Awareness Programme. Master's Thesis, Department of Information Systems and Quantitative Analysis, University Of Nebraska, Omaha.
- [20] Yngström, L and Jörck, F. (2010). The Value and Assessment of IS Education and Training. Department of Computer and Systems Sciences, Stockholm University and Royal Institute of Technology, Electrum 230, SE-164 40 Kista, Sweden. Available Online: <http://people.dsv.su.se/~bjorck/files/infosec-education.pdf>. August 2, 2012.

MANAGING PROJECTS IN A DIGITAL GOVERNMENT

¹B. C. Onyemaobi¹ and V. E. Ejiofor²

¹*Department of Computer Science, Salem University Lokoja*

²*Department of Computer Science, Nnamdi Azikiwe University, Awka*

¹conyemaobi@yahoo.com; ²virguche2004@yahoo.com

ABSTRACT

The process of developing a new system in a Digital Government is a kind of planned transformational change. The digital government projects management ensures the application of information technology knowledge, skills, tools, and techniques to achieve specific targets within specified budget and time constraints. This paper unveils the level of risks inherent in digital government projects and proffers solutions to them. As a consequence of these inherent risks in digital project implementation, there are very high failure rates among digital government applications and government process reengineering projects. Projects related to mergers and acquisitions have similar failure rate. Critical Success Factors (CSFs) and Enterprise analysis methodologies were employed to reduce the inherent risks. These effective management and technical support are required for the success of large-scale government projects. All digital government projects supported with these methodologies are completed on time, on budget, and with all features and functions originally specified.

Keywords: Digital Government, System Development, Project Management, Project Risk, Gantt and PERT charts.

1. INTRODUCTION

Project Management is a carefully planned and organized effort to accomplish a specific (and usually) one-time objective, for example, construct a building or implement a major new computer system. Project management includes developing a project plan, which includes defining and confirming the project goals and objectives, identifying tasks and how goals will be achieved, quantifying the resources needed, and determining budgets and timelines for completion. It also includes managing the implementation of the project plan, along with operating regular 'controls' to ensure that there is accurate and objective information on 'performance' relative to the plan, and the mechanisms to implement recovery actions where necessary. Projects usually follow major phases or stages (with various titles for these), including feasibility, definition, project planning, implementation, evaluation and support/maintenance. Program planning is usually of a broader scope than project planning.

David and Roland (2006), defined Project management as the discipline of planning, organizing and managing resources to bring about the successful completion of specific project goals and objectives. According to Dennis (2007), Project management in the modern sense began in the early 1960s, although it has its roots much further back in the latter years of the 19th century. The need for project management was driven by governments of advanced countries that realized the benefits of organizing work around projects and the critical need to communicate and co-ordinate work across departments and professions. One of the first major uses of project management as we know it today was to manage the United States space programme. The government, military and corporate world have now adopted this practice. Several related project can be managed together, often with the intention of improving an organization's or a nation's workforce performance. In practice and in its aims, project management is often closely related to Information Systems

engineering (Laudon & Laudon, 2014). Information systems are transforming governments and governance. Visible results of this include the increased use of cell phones and wireless telecommunications devices, a massive shift toward online news and information, booming e-commerce and Internet advertising, and new federal security and accounting laws that address issues raised by the exponential growth of digital information. The Internet has also drastically reduced the costs of governance on a global scale. These changes have led to the emergence of the Digital Government (DG). The current digital government projects have usually been focused on the introduction of IT to improve the quality of data and to foster horizontal and vertical integration of back-office and front-office systems. Through this approach, governments are seeking efficiency, effectiveness, and data quality improvement gains, all of them representing a complex pool of governmental and technological challenges. This phase of digital government projects development characterizes most of the current strategies in the developing countries that lead to increased digital projects failure. However, we uphold, like other authors that a digital government projects front end by itself is not a necessary and sufficient mechanism of change. In fact, a more reflective and critical use of IT is required, meaning that government's core processes and their supporting activities should also be considered when developing the digital front end. Thus, to foster an integrated approach of the problem's various edges, a multidimensional model would be a more useful approach.

Project management uses various tools to measure accomplishments and track project tasks and risks involved. These include Work Breakdown Structures (WBS), Gantt charts and PERT (Programme Evaluation and Review Technique) charts. Projects frequently need resources on an ad-hoc basis as opposed to civil services that have only

dedicated full-time positions. Proper Digital Project management reduces risks and increases the chance of success (Dennis, 2007).

1.1 Project Management Triangle

Like any human undertaking, projects need to be performed and delivered under certain constraints. Chatfield and Timothy (2010) stated that traditionally, these constraints have been listed as "scope," "schedule," and "cost". These are also referred to as the "Project Management Triangle," where each side represents a constraint as shown in fig 1. One side of the triangle cannot be changed without affecting the others. The schedule constraint refers to the amount of time available to complete a project. The cost constraint refers to the budgeted amount available for the project. The scope constraint refers to what must be done to produce the project's end result. These three constraints are often competing constraints: increased scope typically means increased time and increased cost, a tight schedule constraint could mean increased costs and reduced scope, and a tight budget could mean increased time and reduced scope.



Fig 1: Project management triangle.
Source: Chatfield and Timothy (2010)

More recently, a further refinement of the constraints in the project management triangle separates product "quality" or "performance" from scope, and turns quality into a fourth constraint as shown in figure 2. This has given birth to the project management diamond, with "time", "cost", "scope" and quality as the four vertices and citizens' expectations as a central theme. No two citizens' expectations are exactly the same, so you must ask what their

expectations are. This research x-rays the discipline and architectural design of Project Management in a Digital Government. It further proffers solutions, tools and techniques that would enable the project team (not just the project managers and governments) to organize their work to meet these constraints.

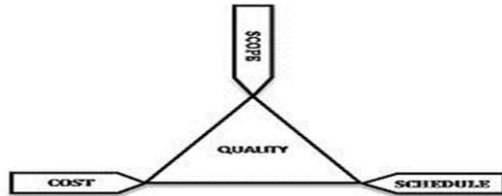


Fig 2: Project management diamond.

Source: Chatfield and Timothy (2010)

)

2. RELATED WORKS

Many projects are being developed and various approaches have been proposed for the design of architectures to deliver digital government services. Few examples include, *egov* (*egov*), *eu-publi.com* (*eupubli*), *egovsm* (mugellini et al, 2005) propose solutions supporting service based systems. However, none of these designs adopts the Critical Success Factors (CSFs) for the representation of concepts and actions.

In the digital government scenario, efforts are under way in which CSFs are involved. The *e-POWER* project (Van Engers 2002) adopts solutions to model inferences, like consistency checking and enforcement in legislation. The *SmartGov* project (SmartGov) developed a knowledge-based platform for assisting public sector employees to generate on-line transaction services. Methodology and Tools for Building Intelligent Collaboration and Transaction Environment in Public Administration (*ICTE-PAN*) proposes a methodology for modeling Public Administration (PA) operations, and tools to transform models into design specifications for public portals (ICTE). ICTE-PAN projects have been initiated to address the G2G (Government to Government) collaboration needs of POs (Public Organizations). Such

projects demonstrated the feasibility of CSFs, although no one explored the opportunities offered by a Semantic Web Service (SWS) infrastructure for the interoperability and integration of services. The *ONTOGOV* project (Onto Gov) develops a platform that facilitates the consistent composition, reconfiguration and evolution of services. It relies upon the SWS technology, although its focus is rather on the service life-cycle than the interoperability and integration issues. ICTE-PAN developed an innovative technology for modeling public administration operations and tools for transforming these models into design specifications for e-Government environments automating and simulating complex bureaucratic processes. Furthermore, meta-tools and peripheral software components have been developed for implementing the design specifications into interactive and intelligent web-enabled portal environments that improve user access to information and facilitate contacts, exchanges and feedback within administrations (ICTE). These imperatives are not new, but many of the solutions are. We can use these modern tools and technologies to seize the digital opportunity and fundamentally change how the Government serves both its internal and external customers – building a 21st century platform to better serve the citizens.

2.1 Phases of Digital Projects

A digital project goes through six phases during its life:

- a. Project Definition: Defining the goals, objectives and critical success factors for the project.
- b. Project Initiation: Everything that is needed to set-up the project before work can start.
- c. Project Planning: Detailed plans of how the work will be carried out including time, cost and resource estimates.

- d. Project Execution: Doing the work to deliver the product, service or desired outcome.
 - e. Project Monitoring and Control: Ensuring that a project stays on track and taking corrective action to ensure it does.
 - f. Project Closure: Formal acceptance of the deliverables and disbanding of all the elements that were required to run the project (Laudon & Laudon, 2014).
- d. Influencing
 - e. Negotiation
 - f. Conflict Management
 - g. Planning
 - h. Contract management
 - i. Estimating
 - j. Problem solving
 - k. Creative thinking
 - l. Time Management (Dennis, 2007).

The role of the Digital Government project manager is one of great responsibility. It is the project manager's job to direct, supervise and control the project from beginning to end. Project managers should not carry out project work, managing the project is enough. Here are some of the activities that must be undertaken:

- a. The Digital Government project manager must define the project, reduce it to a set of manageable tasks, obtain appropriate resources and build a team to perform the work.
- b. The project manager must set the final goal for the project and motivate his/her team to complete the project on time.
- c. The project manager must inform all stakeholders of progress on a regular basis.
- d. The project manager must assess and monitor risks to the project and mitigate them.

No project ever goes exactly as planned, so Digital Government project managers must learn to adapt to and manage transformational changes (Laudon & Laudon, 2014).

A project manager must have a range of skills including:

- a. Leadership
- b. People management (citizens, investors, contractors, suppliers, functional managers and project team)
- c. Effective Communication (verbal and written)

Programme and Project managers bear ultimate responsibility for making things happen in a Digital Government. Traditionally, they have carried out these roles as mere implementers. To do their jobs they needed to have basic administrative and technical competencies. Today they play far broader roles. In addition to the traditional skills, they need to have business skills, citizens relations skills, and political skills. Psychologically, they must be results-oriented self-starters with a high tolerance for ambiguity, because little is clear-cut in today's tumultuous business and political environments. Shortcomings in any of these areas can lead to projects failures.

Many things can go wrong in digital government project management. These things are often called barriers. Here are some possible barriers:

- a. Poor communication
- b. Disagreement
- c. Misunderstandings
- d. Bad weather
- e. Union strikes
- f. Personality conflicts
- g. Poor management
- h. Poorly defined goals and objectives

3. DESIGN METHODOLOGY

There are far more ideas for systems projects than there are resources, hence it is important to select projects that promise the greatest benefit to the Government and complement Government's policies and strategies. Some projects are feasible, while

others may not be feasible for various reasons.

In a developing economy having a Digital Government, the management structure for information systems projects as shown in fig 3 helps ensure that the most important projects are given priority.

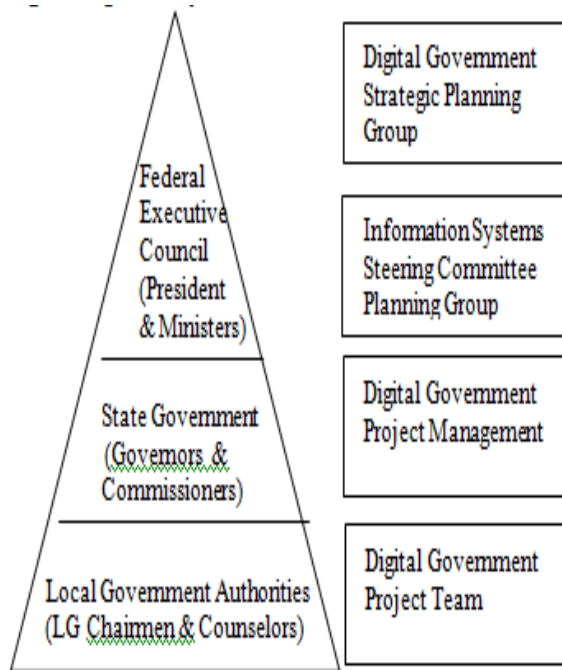


Fig 3: DGcontrol of systemsprojects
 Source: Authors' own conceptualization based on (Laudon & Laudon, 2014)

Each level of government in the hierarchy is responsible for specific aspects of systems projects, and this structure helps give priority to the most important projects for the government.

An information systems plan helps identify projects that will deliver the most strategic value. The plan is a road map indicating the direction of systems development (the purpose of the plan), the rationale, the current systems/situation, new developments to consider, the government strategy, the implementation plan, and the budget. Other important components of an information systems plan include target dates and milestones to help evaluate the plan's future progress and government decisions

regarding infrastructure and transformational changes.

Here we propose two principal methodologies for establishing the essential information requirements of the Government as a wholenamely:

a. **Enterprise analysis(or government systems planning):** Examines the entire government in terms of structural units, functions, processes, and data elements to identify the key entities and attributes of the government's data. The central method used in the enterprise analysis approach is to take a large sample of government officials and ask them how they acquire and use information, what their objectives are, how they make decisions, and what their data needs are. The weakness of enterprise analysis is that it produces an enormous amount of data that is expensive to collect and difficult to analyze.

b. **Critical success factors (or strategic analysis):** This methodology sees the government's success as based on a small number of critical success factors (CSFs) defined by top government officials like Minister of Finance as depicted in fig 4. Digital Government information systems requirements should be focused on providing the information to help meet CSF goals. The principal method used in CSF analysis is personal interviews with three or four top government officials identifying their goals and the resulting national CSFs. Systems are built to deliver on these CSFs. Although this method produces less data than enterprise analysis, there is no particularly rigorous way in which individual CSFs can be aggregated into a clear government pattern, and difficulty may arise distinguishing between individual and national CSFs. The CSF approach relies on interviews with key government officials to identify their CSFs. Individual CSFs are

aggregated to develop CSFs for the entire government.

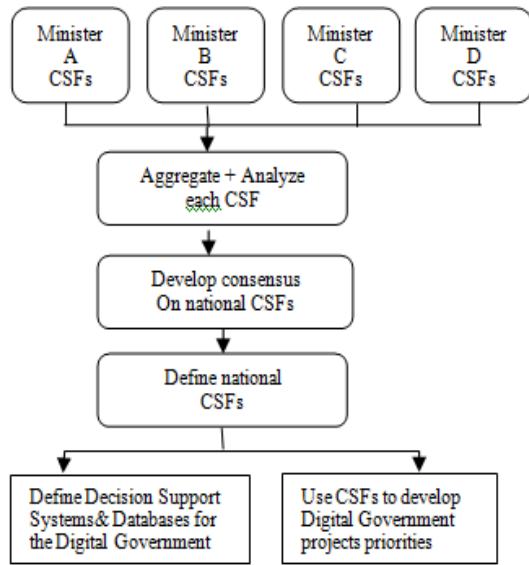


Fig 4: Using CSFs to develop projects priorities
 Source: Authors' own conceptualization

Systems can then be built to deliver on these CSFs.

Once the overall direction of national development has been defined, **portfolio analysis** (an analysis of potential applications within a government to determine the risks and benefits, and to select among alternatives) can be used to evaluate alternative projects. Portfolio analysis inventories all of the national projects and assets, and identifies risks and benefits associated with them. Most desirable, of course, are projects with high benefit and low risk. By using portfolio analysis, ministers can determine the optimal mix of investment risk and reward for their ministries, balancing riskier high-reward projects with safer lower-reward ones. Furthermore, **scoring model** which is a quick method of deciding among alternative systems based on a system of ratings for selected objectives should be used. Scoring models give alternative projects a single score based on the extent to which they meet the selected objectives.

Cost Benefit Analysis (CBA) in Digital Projects

In a way, CBA is the starting point of any system analysis. It is the procedure by which the worthwhile of a system approach is determined, without which there would be no effective idea of the costs and consequences of the new projects being contemplated.

CBA is done on a project that is already underway, there may be pressure to compare all costs and benefits from the beginning of the project. In that situation, the question to be answered is whether or not the benefits of proceeding justify the costs associated with continuing the project. The classic example of this is a situation where large amounts of money have been spent designing a system that has not been successfully implemented, and the project is being re-evaluated. The fact that a lot of money has been spent is no reason to continue spending. CBAs focus on the future and decisions have to be based on the expected costs and benefits of the proposed alternatives. Past experience is relevant only in helping to estimate the value of future benefits and costs.

When Is A CBA Required?

A CBA is always required before a decision is made to initiate, modify or continue a project; the only issue is the level of detail required for the analysis. The process described here is appropriate for a very large, complex, and costly digital project. Scaled down versions of the CBA would be appropriate for smaller, less costly projects; and the government should provide guidelines to determine the amount of scaling that would be appropriate for digital projects based on their size, cost, and complexity.

The CBA is a key input for the investment review that should take place before a new project proceeds to the acquisition or development phase.

The Cost-Benefit Analysis may have to be updated several times during the life cycle

of a system. The first cut at a CBA may be quite brief, and can be used to get concept approval to proceed with a detailed CBA. After the detailed CBA has been completed, the development and implementation plans may call for a prototype system or a pilot phase to test the costs and benefits on a limited scale before the full system is implemented for all users. If that occurs, a third version of the CBA would reflect revised costs and benefits, and would be used to decide whether or not to proceed with full implementation of the project. The post-implementation review of a project may also require an updated CBA to determine if the expected benefits are being achieved, and to decide if the operation should continue as implemented, or if the project should be modified to achieve benefits for the nation to justify continued operation.

Who Should Do The CBA?

One person should be responsible for ensuring that a CBA is done. However, that person will need to assemble a team with expertise in IT systems development and operation, budget, finance, statistics, procurement, IT architecture and the work process being analyzed. A team brings different perspectives to the analysis and the process of estimating costs and benefits, and should ensure more realistic estimates than those of just one person. Additionally, one person rarely has expertise in all of the areas required for a CBA and the knowledge of the work process that is being automated.

4. MANAGING PROJECT RISK IN DIGITAL GOVERNMENTS

Risk Management (RM) is the process of assessing risk, taking steps to reduce risk to an acceptable level and maintaining that level of risk. It also refers to the process of accepting, transferring, or mitigating risk (Laudon & Laudon, 2014).

The level of risk inherent in a DG systems project is influenced by three main dimensions:

- a. **Project size:** The larger or more complex the project, the greater the risk. There are few reliable techniques for estimating the time and cost to develop large-scale information systems.
- b. **Project structure:** Highly structured projects carry lower risk than those with relatively undefined, fluid, and constantly changing requirements; with outputs that cannot be fixed easily because they are subject to citizens' changing ideas; or with populace who cannot agree on what they want.
- c. **Experience with technology:** The project risk rises if the project team and the system project staff lack the required technical expertise.

System implementation generally benefits from high levels of citizens involvement and government support. High levels of populace involvement create opportunities for a system better reflecting national needs, as well as positive reactions to the system by citizens. An important consideration is the discrepancy between the technical orientation of system designers and the political orientation of end-users—a phenomenon known as the user-designer communications gap.

Various project management, requirements gathering, and planning methodologies have been developed for specific categories of implementation problems (Laudon & Laudon, 2014).

- a. **Internal integration tools:** Projects using a complex new technology are riskier and require internal integration tools to ensure that the implementation team is a cohesive unit.
- b. **Formal planning and control tools:** Tools for documenting and monitoring projects that benefit highly structured projects: Gantt charts list project activities, start and end dates of tasks, and

visually represent the timing and duration of different tasks. A project evaluation and review technique (PERT) chart graphically depicts project tasks and their interrelationships. The PERT chart in fig 6 lists the specific activities that make up a web design and hosting project and the activities that must be completed before a specific activity can start. Both Gantt and PERT charts help managers identify bottlenecks and determine the impact that problems will have on project completion times.

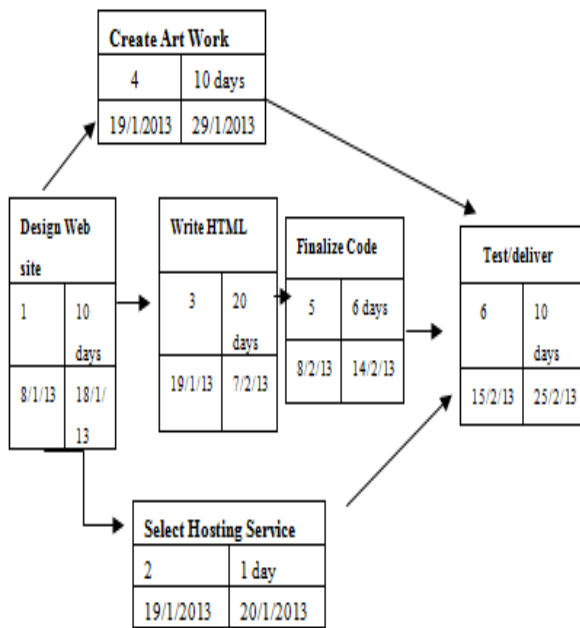


Fig. 5:A PERT chart
 Source: Onyemaobi B. C (2011)

This is a simplified PERT Chart for creating a Website. It shows the ordering of a digital project tasks and the relationship of a task with preceding and succeeding tasks.

c. **External integration tools:** Consist of ways to link the work of the implementation team to users at all levels of government, helping more unstructured projects that need high levels of user involvement. Users can become leaders or active members of the system-building team or they can take charge of training,

installation, and system assessment. Implementation strategy should both encourage user participation and involvement and address the issue of **counter implementation:** a deliberate strategy to undermine the implementation of a digital system. New systems should also address ways in which the digital government will change. **Ergonomics**—the interaction of people and machines in the work environment including such issues as design of jobs, health issues, and the information system end-user interface—must be considered. Systems analysis and design activities should include a **national impact analysis** that explains how a proposed system will affect national structure, attitudes, decision making, and operations.

A **socio-technical design** approach emphasizes participation by the individuals most affected by the new system. The design plan establishes human objectives for the system that lead to increased job satisfaction. Designers set forth separate sets of technical and social design solutions. The social design plans explore different work group structures, allocation of tasks, and the design of individual jobs. The design solution that best fulfils both technical and social objectives is selected. Commercial software tools that automate many aspects of project management facilitate the project management process. Project management software typically features capabilities for defining and ordering tasks, assigning resources to tasks, establishing starting and ending dates to tasks, tracking progress, and facilitating modifications to tasks and resources, and creating Gantt and PERT charts. Microsoft Project has become the most widely used project management software today.

5. RESULTS AND FINDINGS

Digital Government Systems development projects run a very high risk of failure when there is a pronounced gap between users and

technical specialists and when these groups continue to pursue different goals. In addition, if a digital government systems project has the backing and commitment of ministers at various sectors, it is more likely to be perceived positively by both users and the technical services staff. As a consequence of these inherent risks, there are very high failure rates among Digital Government applications and government process reengineering (GPR) projects.

All technology investment projects are completed on time, on budget, and with all features and functions originally specified, with few percent of all digital projects considered "runaway" projects, far exceeding schedules and budgets. In a Digital Government, systems development projects without proper management will most likely suffer these consequences:

- a. Costs that greatly exceed budgets
- b. Unexpected time slippage
- c. Technical performance that is less than expected
- d. Failure to obtain anticipated benefits

Other types of project failings include:

- Systems not being used as intended
- Failing to meet citizens' and government's expectations
- Poor user interface
- Poor data quality

Without proper management, a digital project takes longer to complete and most often exceeds the allocated budget. The resulting system most likely is technically inferior and may not be able to demonstrate any benefits to the government. Great ideas for systems often flounder on the rocks of implementation.

Managing systems development projects in a Digital Government deals with five major variables:

- **Scope:** Defines what work is or is not included in a project
- **Time:** The amount of time required to complete the project.
- **Cost:** Based on the time to complete a project multiplied by the cost of human resources required to complete the project.
- **Quality:** How well the end result of a project satisfies the objectives specified by government.
- **Risk:** Refers to potential problems that would threaten the success of the project.

6. CONCLUSION

Often people underestimate the amount of time needed to implement projects. This is true particularly when the project manager is not familiar with the task to be carried out. Unexpected events or unscheduled high priority work may not be taken into account. Digital government project managers also often simply fail to allow for the full complexity or potential errors involved with a project.

To assuage such lapses the project managers are to make sure that they also allow time for project management administration through CSFs, detailed project, liaison with outside bodies, resources and authorities, meetings, quality assurance, developing supporting documentation or procedures necessary, and training.

7. RECOMMENDATIONS

The digital project manger should allow time for:

- a. Other high urgency tasks to be carried out which will have priority over the present task.
- b. Contact with other citizens, suppliers and contractors.
- c. Others priorities and schedules e.g. local government planning processes.
- d. Quality control rejections.
- e. Unanticipated events (e.g. developing a dynamic website for an institution of

governance and finding that they lack the necessary inputs to keep it up and running).

These factors may significantly lengthen the time and cost needed to complete a project.

If the accuracy of time estimates is critical, you will find it effective to develop a systematic approach to including these factors. If possible, base this on past experience. In the absence of your own past experience, ask someone who has already done the task or project to advise what can go wrong; what you need to plan for; how much and how long each task took previously.

8. REFERENCES

1. Chatfield, C.S and Timothy, J.D.(2010).*Microsoft Project 2010 step-by-step*. Amazon, USA.
2. David, I.C and Roland, G., (2006). *Global Project management handbook*. McGraw-Hill Professional, USA.
3. Dennis L. (2007). *Project management (9rev ed.)*, Gower Publishing, Ltd., Burlington.
4. (eGov) “The eGOV Project Website”. Available on-line: <http://www.egov4dev.org/links/> accessed April 11 2016.
5. (EuPubli) “The EU-PUBLI.con Project Website”. Available on-line: <http://www.eu-publi.com> accessed April 11 2016.
6. (ICTE) “ICTE-PAN Project Website”. Available on-line: <http://www.eurodyn.com/icte-pan> accessed April 7 2016.
7. Laudon, K.C and Laudon, J.P.(2014). *Management Information Systems- Managing the Digital Firm*.Prentice Hall, USA.Pp 567 – 570.
8. Lewis, R. I (2006) *Project Management*.McGraw-Hill Professional.p.110.
9. Mugellini E, Pettenati MC, Khaled OA, Pirri F (2005) eGovernment Service Marketplace: Architecture and Implementation. E-Government: Towards Electronic Democracy. Springer, Berlin. Pp. 193 –204.
10. (OntoGov) “OntoGov Project Website”. Available on-line: https://www.researchgate.net/publication/220082656_Change_management_in_e-government_OntoGov_case_study accessed April 11 2016.
11. Onyemaobi, B.C., (2011). *Managing projects in a digital Firm*.Nigeria Computer Society kogi State Chapter Annual Conference, LokojaNigeria.
12. (OpenCyc) “OpenCyc Project”. Available on-line: <http://www.opencyc.org/> accessed April 11 2016.
13. (SmartGov) “SmartGov Project Website”. Available on-line: <http://www.smartgov-project.org> accessed April 11 2016.
14. Van Engers T., Ptries J.M., Kordelaar J., Den Hartog J., Glassee E. (2002). “ePOWER Project” Available on-line: <http://lri.jur.uva.nl/epower/> accessed April 5 2016.

FUZZY CLUSTER MEANS ALGORITHM FOR THE DIAGNOSIS OF CONFUSABLE DISEASE

G. G. James¹ and A.E Okpako² and J.N. Ndunagu³

¹ *Ebonyi State University, Ebonyi State*

² *Edwin Clark University, Kiagbodo, Delta State.*

³ *National Open University of Nigeria*

¹ *gabresearch@gmail.com*

² *okpako.ejaita@gmail.com, lopito2013@yahoo.com*

³ *jndunagu@noun.edu.ng*

ABSTRACT

Medical science have been overwhelmed in recent times by uncertainty of one form or the other which have greatly affected the decision making process and as such led to cases of misdiagnosis and in worst cases death. One of such forms of uncertainty is the confusability of symptomatic presentations of diseases due to the fact that they share common symptoms and as such becomes difficult for physician's to correctly diagnose them. This difficulty in diagnosis stems from the inability of physicians to quantify the amount of each disease in the confusable disease set depicted by the symptoms. The ultimate goal of medical science is good diagnosis and prevention of diseases and such it is imperative to implement a system to reduce such cases of misdiagnosis which could arise from confusability of disease symptomatic presentations. In this work an expert system driven by the fuzzy cluster means (FCM) algorithms is proposed. The system accepts symptoms as input and provides the degree of membership of each disease in any confusable disease set. Data on alcoholic liver disease were collected and used in the development of the knowledge base. Fuzzy logic and FCM algorithm propelled the inference engine. The system was implemented with CLIPS expert system shell and Java as the front end platform while Microsoft Access was used as the database application. The system gives a measure of each disease within a set of confusable disease. The proposed system had a classification accuracy of 60%.

Keywords: Artificial Intelligence, expert system Fuzzy cluster – means Algorithm, physician, Diagnosis

1.0 INTRODUCTION

Recent advances in emerging computing technologies and communications has posed the question of its suitability in its application to medical sciences. The medical sciences have continuously witnessed a high deluge of data and information due to digital devices that have been infused in medical domain in recent times; yet as good as this may sound, there is need to utilize such data and

information for medical diagnostic procedures since uncertainty of one form or the another may have graciously slipped into the medical data. Medical personnel and physicians with their years of experience and knowledge in their fields can make quality decisions amidst such uncertainty; hence it is imperative to equip computing and associated technologies to handle uncertainty which has led to the use of soft computing in medical

decision support systems.

The quest to make machines intelligent and its use in the medical field in form of clinical support systems has been in existence for decades now, and has transcended into the use of soft computing technologies in making medical decisions so as to combat or mitigate the negative effect or impact of uncertainty in the decision-making process. Complex decision-making requires a lot more than computers can offer. There is an exponential amount of data generated daily in the medical domain, thereby opening doors for all forms of uncertainties such as incompleteness of information, inconsistent description of disease symptoms, overlapping disease symptoms, just to mention a few, and has led to difficulties in properly diagnosing diseases in such situations. Medical uncertainty is an inherent phenomenon in medical science; it is what fuels medical research, prompts patients to seek medical attention and stimulates medical intervention, thus it poses challenges in diagnostic decision-making. In recent times, the negative effect of medical uncertainties has attracted attention due to the emerging realities of this period in medical sciences where evidence-based, shared-decision-making and patient-centered care has brought to fore the limitation of scientific knowledge. The effect of uncertainties in the medical domain has been acknowledged by researchers since the 1950's when the sociologist Renee Fox conducted a seminal study documenting how physicians struggle with uncertainty during their trainings. [13] stated that almost all the physicians are confronted during their formative years by the task of learning to diagnose. Central to good diagnosis, is the ability of an experienced physician to know what symptoms or vitals to throw away and what to keep in the diagnostic process.

Artificial intelligence allows computers to learn from experience, recognize patterns in large amounts of data and make complex decisions based on human knowledge and reasoning skills. Artificial Intelligence (AI)

has become an important field of study with a widespread of applications in fields ranging from medicine to agriculture. AI tools have been applied to medical diagnosis. Expert systems have proven to be an effective and efficient way to diagnose disease, the following methods have been applied in the diagnosis of confusable disease, differential diagnosis:

- *Neural Networks
- *Fuzzy Logic

Confusable diseases share common symptoms and as such overlap, thereby leading to difficult, imprecise or incomplete diagnosis by the physicians since the doctors cannot correctly quantify the amount or degree of each of the diseases represented by these common symptoms. This will not allow the doctor to know the disease to be attended to urgently. Most diseases progress through different stages and complicate the classification process since a disease at one stage can be confused with a different disease at another stage. In this work, an expert system propelled by a fuzzy cluster means (FCM) inference engine is proposed for the diagnosis of confusable disease. The system was implemented with CLIPS expert system shell and Java as the front-end platform while Microsoft Access was used as the database application. The system gives a measure of each disease within a set of confusable diseases. The rest of the paper is organized as follows: section two gives a brief background to medical uncertainties, confusable diseases and fuzzy c-mean algorithm. Section three gives the full description and implementation of the system. Section four discusses results and section five concludes the paper with the summary of the research work.

2.0 Background to Confusable Diseases

Confusable diseases share common symptoms which make it difficult for the physician to establish the right diagnosis [19]. Therefore, diagnostic criteria for a particular disease and needed if the target disease may be confused

with other diseases because of shared symptoms [7]. For a diagnosis, the target disease has to be recognized in a pool of confusable disease. The target disease may be recognized in two ways: by recognition of the combination of symptoms of the target disease or by exclusion of confusable disease as the cause of the symptoms [19].

Figure 1, shows the overlapping features of disease d_i and diseases d_{i+1} and d_{i+2} . In Figure 1, there is no overlap and disease d_i can easily be differentiated from disease d_{i+1} and d_{i+2} . In Figure 2, there is an overlap (indicated by *) and the decision has to be made as to whether patients with symptoms in these overlap areas should or should not be considered as having disease d_i . Diagnostic criteria are not definitive, but are continually being adapted in accordance with new insights and as new research data become available [7].

The following problems are inherent in the diagnosis of confusable disease:

- Confusable disease manifests the same symptoms thereby leading to imprecise or incomplete diagnosis by the physician.
- A disease at one stage can manifest similar symptoms with a different disease at another stage.
- Failure to correctly diagnose a confusable disease would lead to a physician giving the wrong treatment to the patient.
- Patients may be suffering from more than one confusable disease.

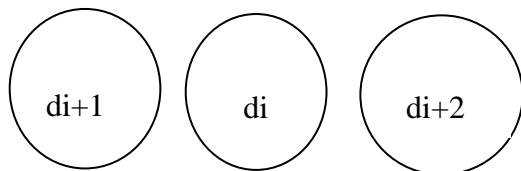


Figure 1: No overlap between disease d_i and disease d_{i+1} and d_{i+2}

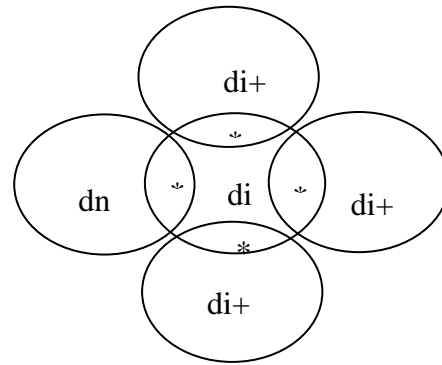


Fig 2: overlapping features of confusable diseases

2.1 AI Tool for Confusable Disease Diagnosis

The different branches of AI have used several tools in the diagnosis of confusable disease. In 2001, Innocent P.R and John R.J proposed a light weight fuzzy process to support early diagnosis of confusable disease using causation and time relationships.

Later on, Innocent and John extended their work that has been applied in the area of clinical diagnosis of confusable disease. They used Fuzzy cognitive maps to encode fuzzy causal structures to aid the diagnosis of confusable diseases.

2.1.2 Fuzzy Logic & Fuzzy Clustering

Logic refers to the study of methods and principles of human reasoning. Classical logic deals with propositions (e.g conclusions or decisions) that are either true or false. Each proposition has an opposite. Thus, classical logic, therefore, deals with combination of variables that represents propositions. As each variable stands for a hypothetical proposition, any combination of them eventually assumes a truth value (either true or false) but never is in between or both (i.e. is not true and false at the same time). The fundamental assumption upon which the classical logic is based is that every proposition is either true or false. This principle has been questioned by many philosophers, ever since Aristotle [15].

Fuzzy logic is a logic whose ultimate goal is to provide foundations for approximate reasoning using imprecise propositions based

on fuzzy set theory, in a way similar to the classical reasoning using precise propositions based on classical set theory. To introduce this notion, we first recall how the classical reasoning works. For instance if say in linguistic terms:

Everyone who is 40 years old or older is old
David is 40 years old and Mary is 39 years old
David is old but Mary is not.

With the above in mind, let's consider the following example of approximate reasoning in linguistic terms that cannot be handled by classical logic:

Everyone who is 40 to 70 years old is old but is very old if he / she is 71 years old or above everyone who is 20 to 39 years old is young but very young if he (she) is 19 years old or below.
David is 40 years old and Mary is 39 years old.
David is old but not very old; Mary is young but not very young.

This is of course a meaningful deductive inference, which indeed has been frequently used in one's daily life. In order to deal with such imprecise inference, fuzzy logic can be employed. Fuzzy logic allows the imprecise linguistic terms such as

- Fuzzy predicates: old, Rare, severe, high, Expensive, Fast.
- Fuzzy Quantifiers: many, few, usually, Much, Almost, Little.
- Fuzzy truth values: very true, true, unlikely true, mostly false, false, and definitely false.

To describe fuzzy logic mathematically, we say that, a fuzzy subset A of a set X is characterized by assigning to each element X being the degree of membership of X in A (e.g. X is a group of people, A the fuzzy set of old people in X).

Two main directions in fuzzy logic have to be distinguished [10]. Fuzzy logic in the broad sense serves mainly as apparatus for fuzzy control, analysis of vagueness in natural

language and several other application domains. It is one of the techniques of soft computing. Fuzzy logic in the narrow sense is symbolic logic with a comparative notion of truth developed fully in the spirit of classical logic.

2.1.2 Clustering

Making sense of data is an ongoing task of researchers and professionals in almost every practical endeavor. Data collection anytime and anywhere has become the reality of our lives. Understanding the data and revealing underlying phenomena are major undertakings pursued in intelligent data analysis (IDA), Data Mining (DM), and system modeling [12].

Categories of Clustering Algorithms

Clustering techniques are diversified; they have been continuously developing for over a half century following a number of trends, depending upon the emerging optimization techniques, main methodology and application domain [12]. The main categories of clustering are:

Hierarchical and Objective function-based clustering [1].

Hierarchical Clustering- produces a graphic representation of the data [14]. The construction of graphs is done in two ways: bottom-up and top-down. In the bottom-up mode, each pattern is treated as a single-element cluster and then the closed clusters are successively merged. At each pass of the algorithm, we merge the two closest clusters. This process is repeated until we get to a single data set or reach a certain predefined threshold value. The top-down approach works in the opposite direction: we start with the entire set treated as a single cluster and keep splitting it into smaller clusters.

Objective Function-Based Clustering is concerned with building partitions (clusters) of data sets on the basis of some performance

index known also as an objective function. The minimization of a certain objective function can be treated as an optimization approach leading to some suboptimal configuration of the clusters [12]. The main design challenge lies in formulating an objective function that is capable of reflecting the nature of the problem so that its minimization reveals a meaningful structure in the data set. Several methods are used to achieve this optimization. The most common one named C-means, is a well-established way of clustering data [20].

Fuzzy Clustering

This class of clustering algorithms allows a pattern to belong to all clusters with different membership grades between 0 and 1. A clustering algorithm that allows for partial membership is regarded as a generalization of the standard fuzzy C-means (FCM). This generalization was introduced by [6] and generalized by [2].

Fuzzy C-Means Clustering

Fuzzy C Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. It is based on the minimization of the following objective, function:

$$J_m^i = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad (1)$$

$$1 \leq m \leq \infty$$

Where m is any real number greater than 1, u_{ij} is the degree of membership of X_i in the clustering j; X_i is the i th element of d-dimensional measured data, c_j is the d-dimension center of the cluster and $\|x_i - c_j\|$ is any norm expressing the similarity between any measured data and center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function with the update of membership u_{ij} and the cluster centers c_j by;

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right]^{\frac{2}{m-1}}} \quad (2)$$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m x_i}{\sum_{i=1}^N u_{ij}^m} \quad (3)$$

This iteration will stop when $\max_{ij} \{ |u_{ij}^{(k+1)} - u_{ij}^k| \} < \xi$, where ξ is a terminal criterion between 0 and 1, whereas k is the iteration steps:

The FCM algorithm is composed of the following steps:

- i. Randomly initialize the membership matrix (u).
- ii. Calculate the centroids c_j .
- iii. Compute dissimilarity between centroids and data points using J_m . Stop if its improvement over previous iteration is below a threshold.
- iv. Compute a new U. return to step 2.

By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the right location within a data set.

2.2 Symptoms of Some Diseases in a Confusable Disease Set

A symptom is a visible or even measurable condition indicating the presence of a disease and hence can be regarded as an aid in diagnosis [3]. A syndrome on the other hand is a collection, a set, or a cluster of concurrent symptoms, which together indicate the presence and the nature of the disease. Joining single symptoms together to one syndrome is one of the main tasks in medical diagnosis.

- Fever
- Jaundice
- Abdominal pain
- Anorexia
- Fatigue
- Ascites
- Splenomegaly

The following are this list of overlapping (shared) symptoms of the two diseases
 The sample symptoms of two diseases in alcoholic liver disease family were selected (i.e. Alcoholic Hepatitis and Cirrhosis). This disease set is given in the table 1.

Table 1: Symptoms of Two Diseases in a Confusable Disease Set

ALCOHOLIC HEPATITIS	ALCOHOLIC CIRRHOSIS
*Fever	*Fever
*Jaundice	*Jaundice
*Abdominal pain	*Abdominal pain
*Anorexia	*Anorexia
	Vomiting

N.B: * indicates overlapping symptoms of this confusable disease set

Fever is considered as a linguistic variable and has the following term set “no fever”, “slight fever”, “high fever”.

Representing this mathematically, we have;
 (Fever) = {“no fever”, “slight fever”, “ high fever”, “very high fever”}

Each linguistic term (e.g. “slight fever”) is associated with a fuzzy set, each of which has a defined membership function (MF). This is known as fuzzifications. The different fuzzification techniques are:

- Triangle MFs
- Trapezoidal MFs
- Gaussian MFs
- Generalized bell MFs

Defuzzification has to be applied to these values to obtain a crisp value that represents this uncertainty. It can be done by applying one of the known defuzzification methods. Such as mean of maximum (MOM) or center of gravity (COG). All the linguistic variables and their term set or linguistic terms are given in Table 2.

Table 2: Symptoms and their Linguistic Values

INDEX	SYMPTOMS	LINGUISTIC VALUES
1.	Fever	“No fever”, “Slight fever”, “High fever”, “Very high fever”.
2.	Jaundice	“No jaundice”, “mild jaundice”, “Deep jaundice”.
3.	Abdominal pain	“No pain”, “mild pain”, “severe pain”.
4.	Anorexia	“Not present”, “present”
5.	Fatigue	“Not present”, “present”.
6..	Ascites	“Not present”, “present”.
7.	Splenomegaly	“Not present”, “present”

3.0 System Architecture

The FCM system architecture adopted in this work is as presented in [4] and is shown in Figure 3.

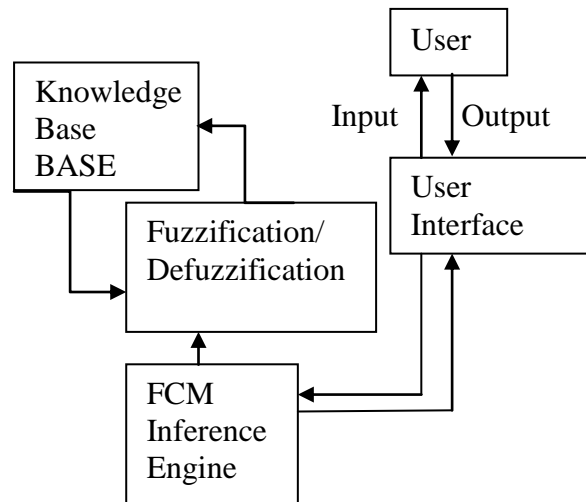


Fig 3: FCM System Architecture (source: [4])

Girratano in [4] describes the main components of an FCM system, being the:

- (a) Knowledge Base
- (b) User Interface, and
- (c) FCM Inference Engine

3.1: The Knowledge Base Sub-Module

The knowledge base stores relevant knowledge needed for the diagnosis. Frames are used in representing knowledge in the knowledge base. Table 3 gives a frame for holding knowledge about alcoholic hepatitis and alcoholic cirrhosis.

Table 3: A Frame for Holding Patient Symptoms HFGHFGHFUFYUY

SLOT NAME	DATA TYPE	VALUE
D _i	Float	*
d _{i+1}	Float	*
...
D _n	Float	*

N.B: Asterisks (*) indicates any float value between 0 and 1

The frame consists of disease as slots. For example, Figure 1 shows a frame for alcoholic cirrhosis.

Figure 1: A frame for alcoholic cirrhosis

Alcoholic Cirrhosis
d _i
d _{i+1}
d _n

The slots d_i ... d_n, where n = 7, holds knowledge about symptoms for alcoholic cirrhosis. For example:

- d₁ holds value for fever
- d₂ holds value for jaundice
- d₃ holds value for abdominal pain
- d₄ holds value for Anorexia
- d₅ holds value for Fatigue
- d₆ holds value for Ascites
- d₇ holds value for Splenomegaly

File Structure

The file structure showed on Table 4 represents the database for storing knowledge in the knowledge base.

Table 4: File Structure

FIELD NAME	DATA TYPE	DESCRIPTION	LENGTH
d _i	Float	Holds value for fever	4
d _{i+1}	Float	Holds value for jaundice	4
...
d _n	Float	Holds value for splenomegaly	4

3.2: User Interface

The user interacts with the system through the user interface. The user enters his/her login information through the user interface, which if correct, displays an interface that allows the user to input patient information and symptoms through the use of check boxes, radio buttons, text fields, etc. The user interface also displays the recommendation on the output panel.

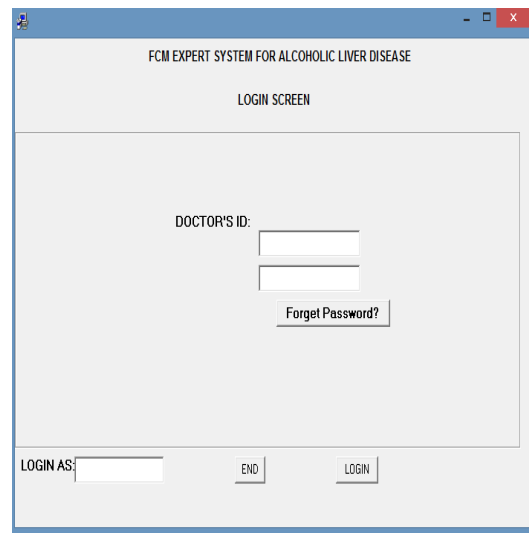


Fig. 4: Login Screen

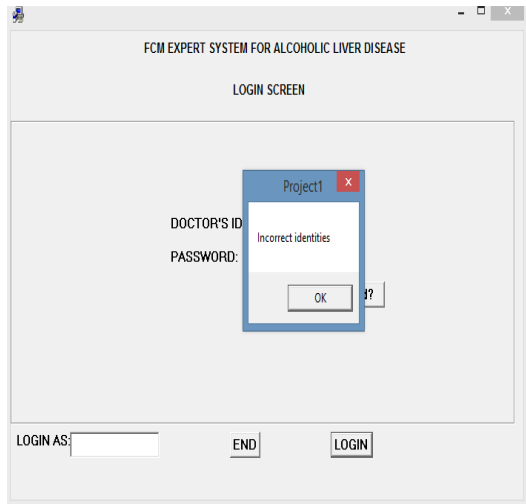


Fig. 5: Login Screen Showing Incorrect Identities

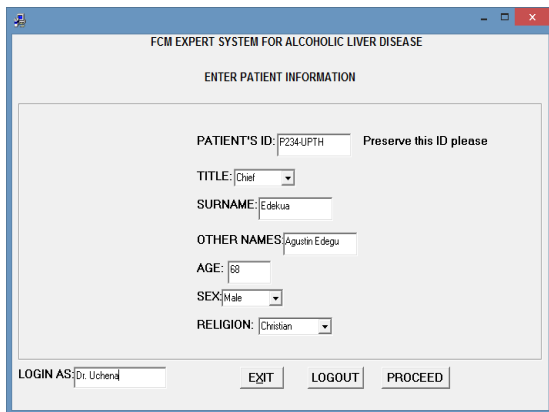


Fig. 6: Patient Information Window

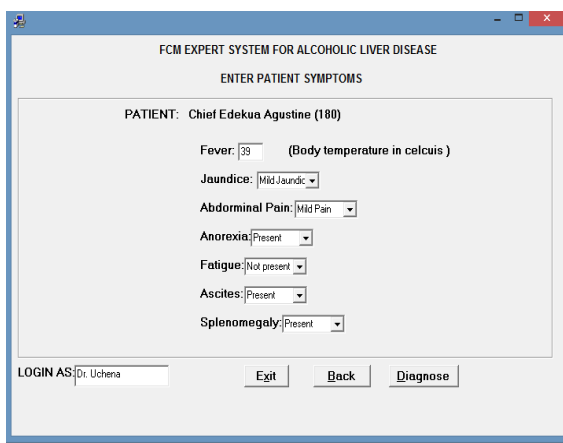


Fig. 7: Diagnosis Window

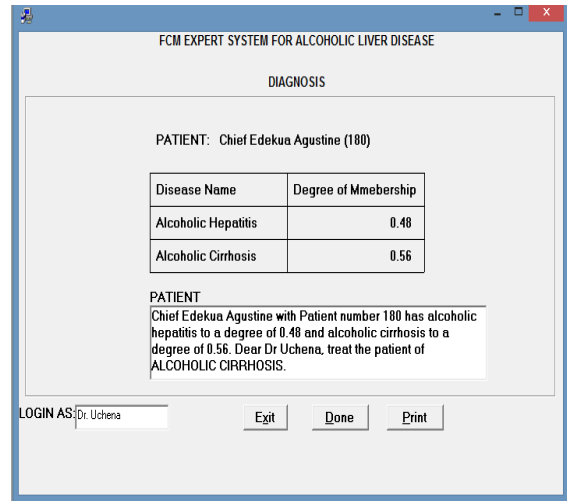


Fig. 8: Recommendation Window

3.3: FCM INFERENCE ENGINE

The inference engine is propelled by the fuzzy cluster means algorithm. The FCM algorithm clusters patient symptoms in this design, the:

- Numbers of clusters, $c = 2$
- Feature vector dimension, $d = 7$
- Fuzziness coefficient, $m = 2$
- Termination criterion, $\epsilon = 0.01$

The FCM algorithm as applicable to this design is composed of the following steps:

STEP 1: Randomly initialize matrix $u = u_{ij}$; Where $i = 1, 2; j = 1, 2, \dots N$.

STEP 2: Calculate the fuzzy cluster centers (centroid) c_j , where

STEP 3: Update $u^{(k)}, u^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^2 \left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right]^2} \quad (4)$$

STEP 4: If $\|U^{(K+1)} - U^K\| < 0.01$, STOP; otherwise return to step2.

MEMBERSHIP FUNCTION GENERATION FEVER

Step 1: Given a data set (body temperature in Celsius).

$X = 38.0, 39.2, 41.5, 36.0, 39.5, 38.6, 40.2, 37.2, 36, 41.0, 41.3, 39.0, 37.0, 38.8,$

40.0, 40.0, 39.4, 39.0, 40.3, 39.9, 36.8, 36.7, 37.5, 39.5, 41.0, 38.4, 39.9, 38.0, 43.0.

The values for X are sorted into ascending order,

X (Asc) = 36.0, 36.4, 36.7, 36.8, 37.0, 37.2, 37.5, 38.0, 38.0, 38.1, 38.4, 38.6, 38.8, 39.0, 39.0, 39.2, 39.4, 39.5, 39.9, 39.9, 40.0, 40.0, 40.2, 40.3, 41.0, 41.0, 41.3, 41.5, 43.0.

STEP 2: The difference between adjacent values in the sorted data is determined which is:

$diff_i = y_{i+1} - y_i = 0.4, 0.3, 0.1, 0.2, 0.2, 0.3, 0.5, 0.2, 0.0, 0.2, 0.2, 0.1,$

Index	8	9	10	11	12	13	14
y_i	39.5	39.5	39.9	40.0	40.0	40.2	40.3
s_i	0.92	1.0	0.98	1.0	0.96	0.98	

0.0, 0.4, 0.0, 0.1, 0.0, 0.2, 0.1, 0.7, 0.0, 0.3, 0.2, 1.5

STEP 3: The similarities between adjacent values were determined using the following formulae:

$$s_i = \left\{ \begin{array}{l} 1 - \frac{diff_i}{5 * 0.28} \\ 0 \end{array} \right\} \dots \quad (5)$$

for $diff \leq 5 * 0.28 - 1.4$

STEP 4: The data is grouped according to similarities using a threshold value α of 0.90.

36.0	36.4	36.7	36.8	37.0	37.2	37.5	38.0	38.0	38.1	38.4	38.6	38.8	39.0	39.0
0.92	0.94	0.98	0.96	0.96	0.94	0.90	1.0	1.0	1.0	0.88	0.96	0.96	1.0	

39.0	39.2	39.4	39.4	39.5	39.5	39.9	40.0	40.0	40.2	40.3	41.0	41.0	41.3	41.5	43.0
0.92	0.94	0.98	0.96	0.96	0.94	0.90	1.0	1.0	1.0	1.0	0.88		0.96	0.96	1.0

This produces four groups.

Table 5: Class 1

Index	1	2	3	4	5	6	7
y_i	36.0	36.4	36.7	36.8	37.0	37.2	37.5
s_i	0.92	0.94	0.98	0.96	0.96	0.94	

Table 6: Class 2

Index	1	2	3	4
y_i	38.0	38.0	38.1	38.4
s_i	1.0	1.0	1.0	

Table 7: Class 3

Index	1	2	3	4	5	6	7
y_i	38.6	38.8	39.0	39.0	39.2	39.4	39.4
s_i	0.96	0.96	1.0	0.96	0.96	0.98	1.0

Table 8: Class 4

Index	1	2	3	4	5
y_i	41.0	41.0	41.3	41.5	43.0
s_i	1.0	0.94	0.96	0.70	

Steps 5: The membership function for each class is derived using the formula:

$$b_j = \frac{y_i * s_i + y_{i+1} * \frac{s_i + s_{i+1}}{2} + A + y_{k-1} * \frac{s_{k-2} + s_{k-1}}{2} + y_k * s_{k-1}}{s_i + \frac{s_i + s_{i+1}}{2} + A + \frac{s_{k-2} + s_{k-1}}{2} + s_{k-1}} \quad \text{--- (6)}$$

$$a_j = b_j - \frac{b_j - y_i}{1 - \mu_j(y_k)} \quad \dots \dots \quad (7)$$

$$\text{similarly, } c_j = b_j - \frac{y_k - b_j}{1 - \mu_j(y_k)} \quad (8)$$

Table 8 shows the value of b_j , a_j , and c_j for each class.

Table 9: End Point of Each Membership Function

	a	B	c
Class 1 (no fever)		37.0	38.5
Class 2 (slight fever)	37.2	38.7	40.2
Class 3 (high fever)	38.2	39.6	41.0
Class 4 (very high)	40.2	42.2	

nf = no fever
 sf = slight fever
 hf = high fever
 vhf = very high fever

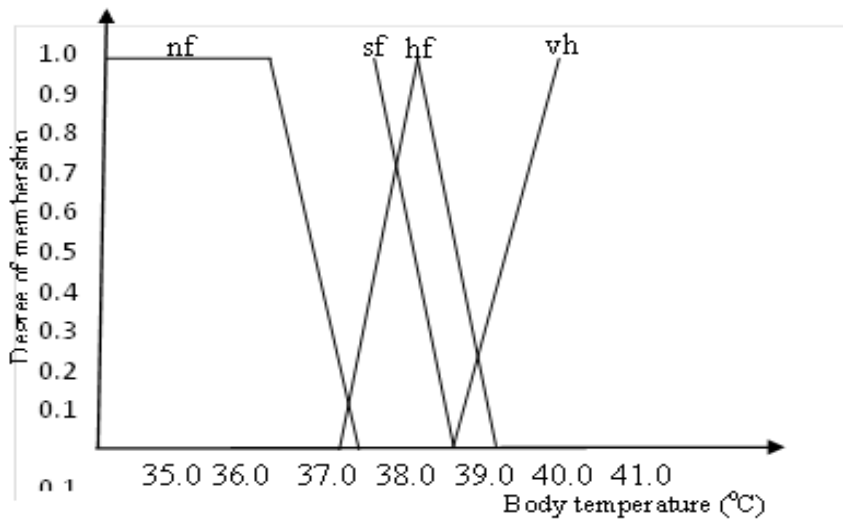


Fig. 4: Graph of membership function to body temperature

$$m_{nf}(x) = \begin{cases} 1 & \text{If } x < 37.0 \\ \frac{38.5 - x}{1.5} & \text{If } x > 38.5 \\ 0 & \text{If } 37.0 < x < 38.5 \end{cases}$$

$$m_{sf}(x) = \begin{cases} 0 & \text{If } x < 37.2 \\ \frac{x - 37.2}{1.5} & \text{If } 37.2 < x < 38.7 \\ \frac{40.2 - x}{1.5} & \text{If } 39.6 < x < 41.0 \\ 0 & \text{If } x \geq 40.2 \end{cases}$$

$$m_{hf}(x) = \begin{cases} 0 & \text{If } x < 38.2 \\ \frac{x - 38.2}{1.4} & \text{If } 38.2 < x < 39.6 \\ \frac{41.0 - x}{1.4} & \text{If } 39.6 < x < 41.0 \\ 0 & \text{If } x \geq 41.0 \end{cases}$$

$$m_{vhf}(x) = \begin{cases} 0 & \text{If } x < 40.2 \\ \frac{x - 40.2}{2.0} & \text{If } 40.2 < x < 42.2 \\ 1 & \text{If } x < 42.2 \end{cases}$$

The grades of membership are then defuzzified using the mean of maxima defuzzification method.

Mean of maxima

$$U = \sum_{i=1}^R u_i / R \dots \dots \dots (9)$$

JAUNDICE

The term set for the linguistic variable jaundice take on the crisp value 0, 1, and 2 respectively, that is:

- Not present = 0
- Mild jaundice = 1
- Deep jaundice = 2

ABDOMINAL PAIN

The term set for linguistic variable abdominal pain take on the crisp values 0, 1, and 2 respectively, that is:

- No pain = 0
- Mild pain = 1
- Severe pain = 2

ANOREXIA, FATIGUE, ASCITES AND SPLENOMEGALY

The term set for the linguistic variables anorexia, fatigue, ascites, and splenomegaly take on the crisp values 0 and 1, respectively, that is:

- Not present = 0
- Present = 1

4.0 Results and Discussion

Fuzzy C-means algorithm is applied to patient data for alcoholic liver diseases collected at University of Uyo Teaching Hospital. Eighteen (18) correct classified samples were obtained out of Thirty (30) samples used for the testing. Table 10 gives the result in a tabular form.

Table 10: Classification Result

Classifier	No. of correct Classified Samples	Total No. of Samples	Correct Classification Rate
Fuzzy C-Means Algorithm	18	30	60%

These result shows that FCM algorithm can be applied to confusable disease diagnosis.

5.0 CONCLUSION

The proposed system would give great assistance to physician as it would help them determine the degree of each disease present in a patient and reduces administration of wrong treatment. Our aim of applying the FCM algorithm in the diagnosis of confusable disease has been realized. As a result, FCM algorithm could be used as an important supportive tool for medical experts in diagnosing confusable diseases.

REFERENCES

- [1]. C. J .Bezdek, R. Krishnampuran and N.R. Ral. Fuzzy models and Algorithms for pattern Recognition, Kluwer Academic Publishers, Aordercht. 1999.
- [2]. C. J. Bezdek, R. Ehrlich and W. Full. FCM: the Fuzzy C-Means Clustering Algorithm. Computers and Geosciences 10: 191-203. 1984.
- [3]. George Becks, M. Dotoli, Diechrich craf key-serlink and Jan Jantzen. fuzzy clustering. A versatile means to Explore medical Database, ESIT Aachen Germany, 2000.
- [4]. Girratano Reley: Expert systems. The Development and implementation of Expert system, mograw Hill Inc. USA, 2005.
- [5]. Hong Tzung-pei and C .Lee. Induction of fuzzy rules and membership functions from Training examples, fuzzy sets and systems. 84, 33-47, 1996.
- [6]. J. C. Dunn. Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions. IEEE Transactions on Computers 24 (4): 835-838. 1975.
- [7]. J. F. Fries, Ann Rheum. Alcoholic liver Disease. Proposed diagnostic criteria for classification, 2004.
- [8]. K .Kalantar-Zadeh , M. Kleiner , E. Dunne, G.H Lee and F.C Luft . A modified quantitative subjective global assessment of nutrition for dialysis patients. Nephrol Dial Transplant. Jul; 14 (7):1732-8. 1999.
- [9]. K. P Adlassening. Fuzzy set theory in medical diagnosis. IEEE Transactions on system, man, cybernetics, smc-16,2, 260-265 1996.
- [10]. L.A. Zadeh et al.. “ Decision theory with imprecise probabilities, ” in Contract on Application of Fuzzy Logic and Soft Computing to Communications, Planning and Management of Uncertainty, Berkeley, Baku, 2009.
- [11]. Maria columeness and Olaf Wolken-Hauer. An Introduction into fuzzy clustering, 2008.
- [12]. P. Witold . Fuzzy Relational Clustering Algorithm. Advances in Fuzzy Clustering and its Applications. A book edited by Jose Valente de Oliveira, Witold Pedrycz. 1999.
- [13]. R. Brause and F. Friedrich. A neuro-fuzzy approach as medical diagnostics interface, In Proceeding of European Symposium on Artificial Neural Networks (ESANN). 201-206, 2000.
- [14] R. Duda , P.Hart and D.Stork. Pattern Classification. 2 edition. New York: John Wiley & Sons. 2001.
- [15]. R. I. John and P.R. Innocent. Modeling uncertainly in clinical

- diagnosis using fuzzy logic, Centre for computational intelligence, 2004.
- [16]. R. O. Huang and C. A. Kulikowski. Computer-Based medical consultation: MYCIN, Newyork, Elsevier, 1999.
- [17]. Richard Robertson and J. H Friedman. A fuzzy system for helping medical Diagnosis of malformations cortical development, 2008.
- [18]. S. Moein, M. L. Gigar, C. J Doik, Vyborny, and R.A Schmidt. A Novel fuzzy Neural based medical diagnosis system, 2008.
- [19]. V. M. Joop. Diagnosis and differential diagnosis of Alcoholic liver Diseases, 2005.
- [20]. A web. Statistical pattern Recognition, 2nd edition, John Wiley, Aoboken, N J, (2002):

A COMPUTERIZED LEGAL INFORMATION MANAGEMENT SYSTEM

***K. Ohiagu¹ and O. Omorogiuwa²**

*^{1,2}Department of Computer Science/Information Technology
College of Natural and Applied Sciences*

Igbinedion University

Okada, Edo State, Nigeria

¹kingsohiagu200@gmail.com, ²ask4osas@yahoo.com

ABSTRACT

In recent times, incidences of social vices in our society has been unprecedented. The deluge of legal cases our courts receive on a daily basis has posed a great challenge to the Judiciary with respect to accurate information management. In response to this, the study set out to investigate the problems associated with the manual record keeping process through the filling system using the survey research methodology. Firstly, the manual record keeping system done via the filling caused problems such as situations where records kept in files are either outrightly missing or misplaced and files not properly placed in their appropriate locations thereby not providing easy access to legal information as at when required for the making of critical judicial decisions. Secondly, the risk involved in unauthorized disclosure of vital information was also adduced. A framework for the design and implementation of a legal information management system was presented. Ultimately to show the functionality of our system, the framework was translated into a software using some Object Oriented Design Tools, a suitable event-driven Programming Language; visual basic 6.0 was adopted and Microsoft Access was used as the database component.

Keywords: Legal Information, Court Case, Judiciary, Manual, Records, event-driven

1.0 BACKGROUND INFORMATION

Proper legal information management is the backbone of an efficient judiciary. There is no gainsaying that the need for accurate recording keeping in the court system cannot be overemphasized. The Judiciary being the last bastion of hope for the common man, if unable to provide the much desired impartial and all-inclusive legal decisions has failed woefully in its primary responsibility as the custodian of legal information. With the growing incidence of social vices in our society, there has been an astronomical rise in legal information. This has posed a challenge to the Judiciary with respect to accurate information management; Courts around the

globe are embracing ICT (Information and Communication Technology) as a veritable tool to help provide the enabling platform for quick, reliable and consistent dispensation of justice.

1.1 RESEARCH PROBLEM

The Federal High Court, Benin typifies a cosmopolitan court setting which should be seen as a repository for legal information. Presently the real time legal information management infrastructure available is manual and fraught with myriad of shortcomings such as files containing records were either out rightly missing or misplaced and files not properly placed in their

appropriate locations thereby not providing easy access to legal information as at when required causing a near total breakdown of court hearing and court judgment procedures and also the risk involved in the unauthorized disclosure of vital legal information to members of the public which could have dire consequences and also the propensity of thwarting a judicial process. Based on the foregoing, there is a dire need to correct this growing anomaly rearing its ugly head within the judiciary.

1.2 RESEARCH DIRECTION

The research is intended to design and implement a computerized legal information management system for the Federal High Court in Benin. This will provide a repository for keeping legal information, provide facilities for search and retrieval of legal information where applicable for the purpose of enquires and investigations, provide an audit trail for keeping track of user logins and legal information inputted by a user at any point in time and finally allow users to generate various legal information reports

2.0 RELATED LITERATURE

Legal information resources are essential ingredients for effective legal research [1] Effective records management system guarantees the accountability and integrity of an organisation that provides services to the public at large and serves as strategic resource for government administration [2] Without a reliable and accurate case file system day-to-day court operations would be hampered and this will in turn affect judicial decisions. The maintenance of case records directly affects the timeliness and integrity of case processing. There is a pressing need for a clear definition of legal framework [3] [4] alleged that in order to minimize the risks and costs of regulatory and legal non-compliance, litigation, discovery, business inefficiency and failure, organizations need to remove the human element by automating records management via technology. This transformation means removing freedom of choice; enforcing electronic record creation,

indexation; classification, naming conventions (thesaurus and taxonomies), creation and preservation of meta-data, minimizing duplicate records by creating a central information repository which will also facilitate knowledge and content management, systematically archiving and tracking records and amendments, applying retention schedules to purge redundant ones, but preserving their access logs, audit trails and meta-data.

[5] and [6] believe that the major issues of electronic records in organizations are regarding access, security and interoperability. According to [7] Interoperability refers to the ability of different Information Technology (IT) systems and software applications to communicate to exchange data among them accurately, effectively and especially to use the information that has been exchanged. According to [8] organization today not only have to comply with regulations, but also have to maintain a balance between operational record keeping requirements, minimizing liability of storing private information and customer privacy preferences. The biggest challenge when organizations set to move forward by embracing IT in its administration is to retrain the people. For a court registry, the lack of experts who know both registry office and information management standards becomes the first hurdle in implementing change. [9] pointed out a number of issues identified by legal and judicial record case studies with respect to people aspect: the need for consistent and authoritative instructions on the preservation or destruction of court case records (both paper and electronic); the importance of having a high level 'champion' within the courts to promote good practice in records and information management; the need for professionally trained records managers within judiciaries; the need for formal training and training materials in judicial records and information management and the importance of having expert advice and guidance available to those with responsibility for records and information

management in the courts.

[10] identified a number of electronic court applications and services being implemented in a few countries e.g. United States of America had the PACER, Australia had the eSearch - for public to search cases, eFiling – electronic document lodgement, eCourtroom - virtual courtroom for pre-trial matters, eCase Administration - for legal practitioners and parties to communicate with court chamber staff securely, Commonwealth Courts Portal, Singapore had - eAlternative Dispute Resolution, eJustice Judges' Corridor, Justice Online (JOL)- a global forum and virtual think-tank for judges to mention a few.

Experts from the United States, Europe, Australia and Singapore, inspired by court quality models used in a number of international communities, formed a Consortium with the goal to take necessary steps to achieve court excellence. The Consortium concluded that the most effective way to achieve this goal was to develop a framework called “International Framework for Court Excellence”. The Framework assesses a court’s performance against seven areas of excellence and provides guidance for courts to improve their performance. It utilizes recognized organizational improvement methodologies while reflecting the special issues that courts face. The International Research on Permanent Authentic Research in Electronic Systems (InterPARES) project based in the University of British Columbia (UBC) brings together archivists from universities and archival institutions, along with computer and information scientists and engineers from around the world in a concerted effort to define the archival requirements for authenticity on the basis of archival science and diplomatics ([8], [3]).

3.0 MATERIALS AND METHOD

To substantiate the need for computerization of legal information in Federal High Courts,

especially the one in Benin and Okada, we designed a questionnaire and conducted personal interviews. The questions consist of five sections; Section A consist of nominal scale questions representations, Sections B and C consist of Ordinal scale questions representation while Sections D and E consist of the use of 5 Likert scale measurement. The questionnaire was distributed in two high courts office complex in Benin and Okada. A total of 30 respondents (eighteen males and twelve females) responded to the questionnaires. Data gathered was analyzed using descriptive statistics. Find attached a sample of the questionnaire in appendix A.

3.1 ANALYSIS OF THE QUESTIONNAIRE

Section A captured the respondents’ basic bio data and their current job placement category. Sections B and C was administered to establish the availability and usability of ICT facilities in the High Court. Table 1.0 showed the descriptive statistics of the respondents’ view to Sections B and C of the questionnaire. The respondents showed a high level of usability of the various facilities within the High Court e.g. 65% can effectively use the computers while approximately 51% of the respondents indicated that they can use the available facilities (i.e. Internet facilities, digital camera, scanner, projector and printers). This shows their readiness to accept the computerization of the existing manual process.

Table 1.0: Respondent view on ICT Availability and Usability in High Court

Table 2.0 showed a description of the respondents view to section D of the questionnaire. 79% of respondents agreed that lack of expertise is a major factor that hindered the use of ICT. Although 70% agreed that they are already conversant with the manual process but 72% disagreed that using computers will reduce the staff strength.

S/N	Facilities	Availability	Usability
1.	Computers	70%	65%
2.	Internet Facilities	40%	50%
3.	Digital Camera	30%	45%
4.	Scanner	60%	60%
5.	Projector	30%	45%
6.	Printers	10%	55%

Table 2.0: Respondents views on ICT facilitate usage in the High Court

S/N	Questions	Strongly Agree (%)	Agree (%)	Undecided (%)	Disagree (%)	Strongly Disagree (%)
1.	Lack of expertise/skills required to operate equipment	79	12	1	6	3
2.	No training in how to use some available equipment	10	13	2	50	25
3.	Lack of confidence in using ICT facilities	7	2	1	30	60
4.	Unavailability of required ICT facilities	4	3	1	25	67
5.	Already used to the manual process.	70	10	3	7	10
6.	The fear that computerizing the legal process will reduce the use of manual labour	6	4	10	22	72

S/N	Questions	Always (%)	Often (%)	Sometimes (%)	Rarely (%)	Never (%)
1.	How often is a new client case registered manually.	95	1	2	2	0
2.	How often are client case files updated manually.	94	2	3	1	0
3.	How often are client case files removed from their storage location.	90	2	4	3	1
4.	How often do client case files get lost	20	10	10	55	5
5.	How often are court summons issued.	90	2	1	5	2
6.	How often are court warranty issued.	80	3	10	6	1

ICT Facilities Usage in the High court

Table 3.0: Respondents view on the use of manual process of carrying out operations in the High Court.

Table 3.0 which is formulated from data gathered from Section E of the questionnaire represents a descriptive statistics of the various respondents view to the use of the present manual processing system. 95% of the respondents showed that manual processing system is the major method of processing information in the High Court. Due to the cumbersome nature of this method of processing information arising from human fatigue, 20% of the respondents agreed that client files sometimes get lost, while 94% said updating client and case files manually are so frequent and stressful. The findings from the respondents suggested the need for this system and willingness of Staff to embrace the use of computers to improve service delivery in the Federal High Courts. We were therefore motivated by this obvious lapse in legal information processing to develop a Computerized Legal Information Management System.

3.1 RESEARCH APPROACH

The Structured System Analysis and Design approach was adopted. Some data gathering techniques was used to elicit some relevant information required to comprehend the existing system. Questionnaires/Personal interviews was used to ascertain the efficacy of the manual legal record keeping process from a randomly selected sample of Judiciary and legal staff of the Federal High Court, Benin and Okada. An aspect of the personal interviews was used to ascertain the state of the manual process with regards to its possession of the required platforms needed to support the proposed system.

3.2 ISSUES ASSOCIATED WITH THE EXISTING SYSTEM

The system we are investigating is the existing manual record keeping system. It is a manual system and fraught with a lot of

shortcomings. From the survey, all (Judiciary & legal staff) keep records manually via the filling system. Based on this, files containing records were either out rightly missing or misplaced and files not properly placed in their appropriate locations thereby not providing easy access to legal information as at when required causing a near total breakdown of court hearing and court judgment procedures.

Another shortcoming with the existing system is the risk involved in the unauthorized disclosure of vital legal information i.e. details of a knotty court judgment for a complicated court case to members of the public. For example, an authorized staff of the court only should have access to this vital information by virtue of the fact that he/she possesses of valid login information into the system if it had been in existence.

3.3. NEED FOR PROPOSED SYSTEM

Technology has provided the enabling platform for prompt and accurate information transfer.

Without a reliable and accurate case file system, day-to-day court operations would be greatly hampered and this will in turn affect judicial decisions. Therefore, the need for the new system cannot be overemphasized. Based on this, conceptual solutions were translated into appropriate designs. The processes and tools used to achieve this is undertaken next in the proposed system design methodology.

4. PROPOSED SYSTEM DESIGN

The modular design approach was used for the proposed system design. The proposed system is a computerized legal information management system. The system is user-friendly. It will be designed with features which provide users with an input

screen which allows a user to input his/her login information to the system. After a user has successfully login into the system, the user can create a record of a client which is stored in a database, the user can update or delete a client's already existing in the database by using three search parameters such as a client's case file number, date client's case file was filed and client's name to query it to ascertain the records to be deleted or updated. The user can view a court case information of a client by using any of the search parameters elucidated above to

query the database for the client's records which are displayed on the screen. In a circumstance where judgement was handed down on a particular client and the information already inputted into the system the user can also query the database using the search parameters to view the court judgement of the client in question. The user can also view a daily case calendar by querying the database using a court date as search criteria The legal information management system architecture below Fig 1

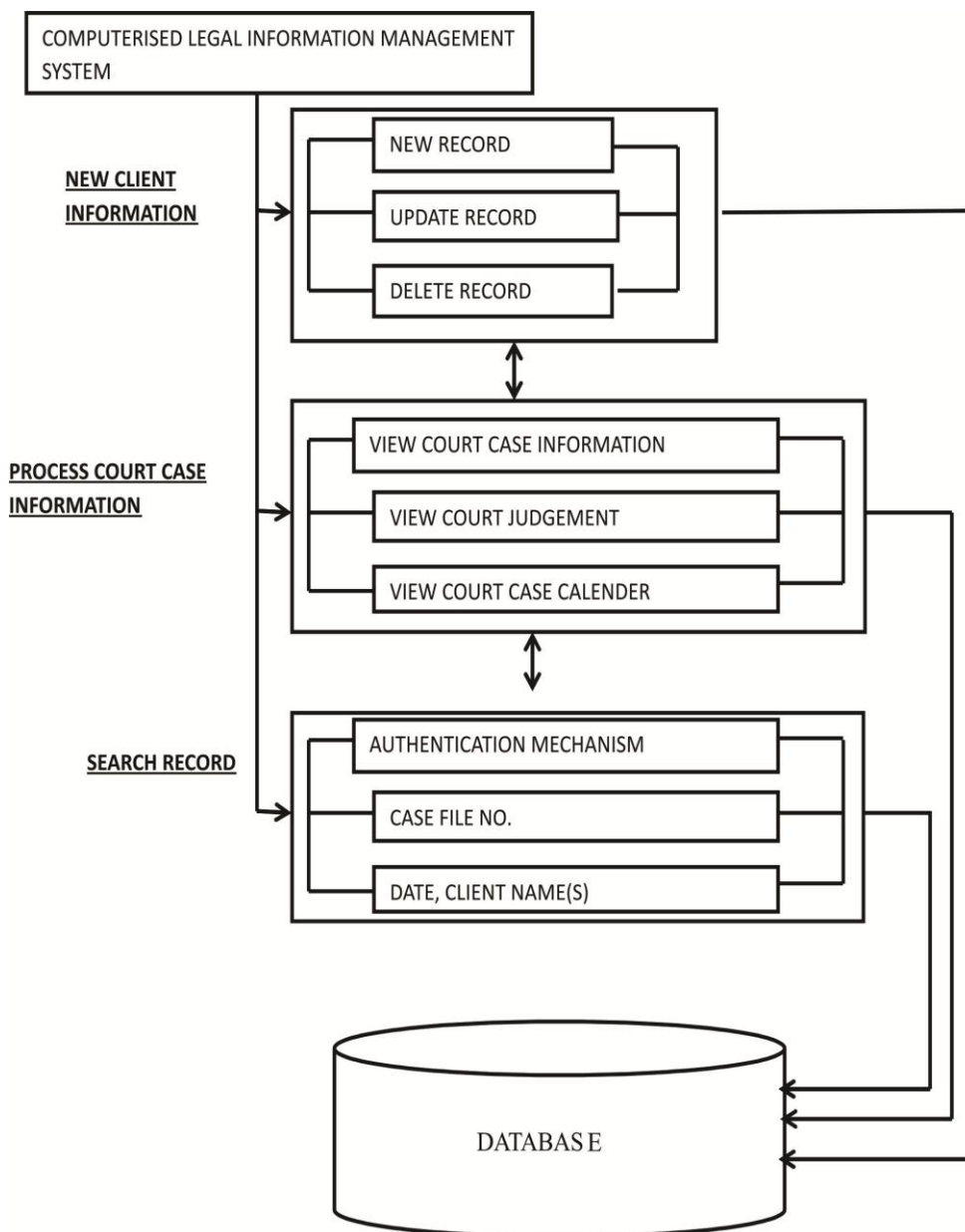


Fig1: Overall System Architecture for the Computerized Legal Information System

4.1 PROGRAM DESIGN

The software was developed in Visual Basic 6.0 Programming Language Integrated Development Environment (IDE) as the front end and Microsoft Access as the back end and designed using the modular approach. The choice of this IDE was informed by its features that supports coding and testing for the required functionalities. The software was then tested on this same platform to gauge the extent to which the desired functionalities were accomplished. There were errors revealed in the course of the testing, with careful design of the software however, errors were minimized. Several modules were fused together to form the entire legal information management system, each module was designed separately and linked together afterwards. There is an in-built security feature where the password approach is adopted to ensure that only authorized users have access to the system information.

5. SYSTEM IMPLEMENTATION

The software can be installed on any system with at least Microsoft Windows 7 operating system or a higher version. To launch the software, click the start button, click on a computerized legal information management system which was added to the start button after successful installation of the software, this will launch the software displaying a splash screen. After a few seconds the user is prompted by the system to provide login information, after successful authentication by the system the main menu is displayed (Fig 2).

To register a new court case information, click on the Register new court case sub menu this will display a new court case entry form where the appropriate details are inputted (Fig 3).

To update a court case information, click on update case records submenu, this will display a search screen where a court case number is inputted to display the case information for that case number and to enable an update to be effected via an update button (Fig 4).

To delete a court case, click on delete case records submenu, this will display a search screen where a case number is inputted to display the case information for that case number to enable the user effect a delete operation via the delete button (Fig 5).

A search is done using three parameters that is by case file no, client's name and date case was filed by client. In the event of wrong search operation an alert message box is displayed on your screen (Fig 6).

To update a court case judgement information, click on the update court case judgement sub-menu, this will display a search screen where a case number is inputted to display the case information for that case number to enable the user effect an update to a court case judgement via an update button (Fig 7).

To display a court case information report, click on court case info submenu, this will display the court case information report screen. The user will be expected to input a court case number to search for the court case information associated with that case number, if found, a report is displayed on the screen (Fig 8).

To display a daily case calendar, click on daily case calendar sub-menu, this will display the screen for the user to input the required court date. The user invokes a search by clicking on the search button, a sample calendar report which matches the user specification is displayed (Fig 9).

*A Computerized Legal Information Management System
K. Ohiagu and O. Omorogiuwa*

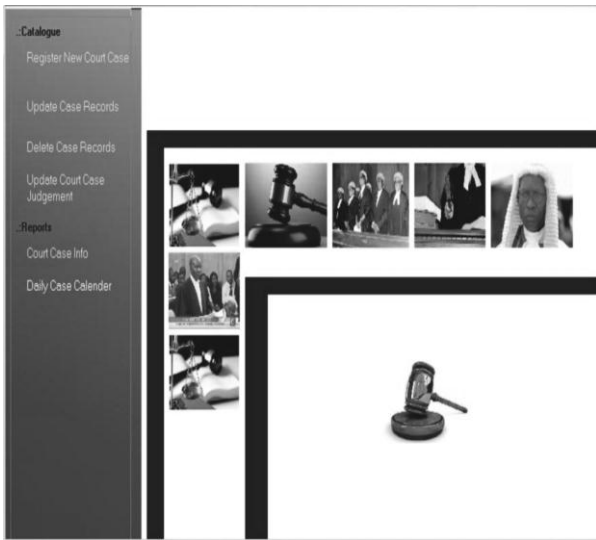


Fig 2: Main menu window of the Computerized legal information management system



Fig. 3: Register a new court case information window



Fig 4: Court case information update window



Fig 5: Court case information delete window

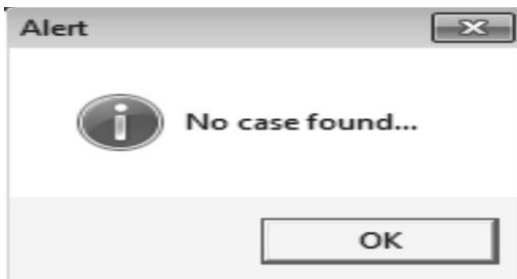


Fig 6: No records found alert message box window



Fig 7: Court case judgement update window



Fig 8: Court case information report window

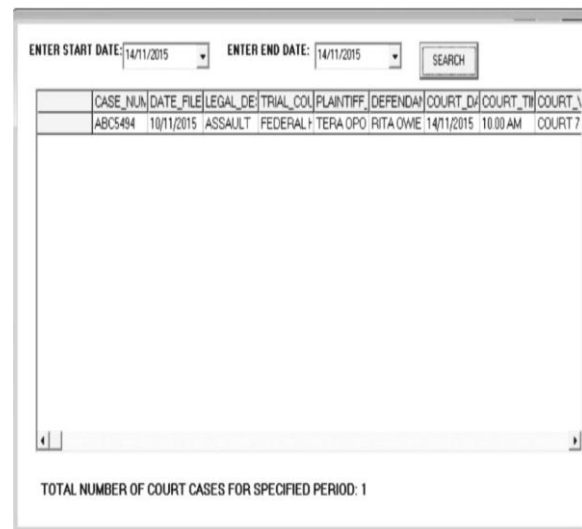


Fig 9: Case Calendar report window

6. CONCLUSION

In this study, a framework for the design and implementation of a Computerized Legal Information Management system for Federal High Court, Benin was presented. Questionnaires/Personal interviews were used to ascertain the efficacy of the manual legal record keeping process from a randomly selected sample of Judiciary and legal staff of the Federal High Court, Benin. The implementation of this system will immensely help to alleviate the sufferings of

judicial staff of the Federal High Court Benin from the existing manual system with respect to prompt processing of legal information which would in turn hasten court proceedings as much as provide quicker and fairer judicial decisions. Once the system is used according to specification with the validations checks in place the system will undoubtedly work perfectly for system users. Further research work in this direction can be done by adding more features to enhance the system further.

REFERENCES

1. Anyaogu and M. Iyabo Legal information resources are essential ingredients for effective legal research. *Information and Knowledge Management Journal*, 4(9): 50-58, (2014)
2. K. Hassan Court Records Management in Malaysia. *Personal Communication*, 3(2): 12-19, (2007).
3. Y. Johare, E-government and Records Management: An Assessment Tool for e-Records Readiness in Government. *The Electronic Library*, 14(2): 56-64, (2007)
4. H. Gouanou. and J. Marsh Electronic Filing System (EFS). *The Electronic Library*, 25(3): 274 – 284, (2014).
5. Z. Manaf, and A. Ismail Malaysian Cultural Heritage at Risk : A Case Study of Digitization Projects. *Library Review*, 59(1): 107-116, (2010)
6. A.Ojo, T. Janowski, and E. Estevez Semantic Interoperability Architecture for Electronic Government Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government, Digital Government Society of North America, pages 63-72, (2009).
7. A. Atallah A Framework for Records Management in Relational Database systems Masters Degree Thesis, University of Waterloo, Ontario, Canada (2010).

8. L. Duranti, From Digital Diplomats to Digital Records Forensics *Archivaria*, 68(1): 39-66. (2009).
9. S.Akinyemi, Electronic Records Management: New Obligations, New Tools *Community Banker*, 3(3):11-17. (2014).
10. W. Saman, and A. Haider, Towards the Court Records Management Reform In Proceedings of the National Conference on Shariah and Law, Kuala Lumpur, November 2-3. (2015).

Appendix A: Research Questionnaire

This questionnaire is designed to establish the need for a Computerized Legal Information Management System in Nigeria. Kindly answer the questions carefully and sincerely by ticking the appropriate space provided.

Section A: BACKGROUND INFORMATION OF RESPONDENTS

1. Gender: Male() Female()
2. Age 21-30() 31-40() 41-50() 51 & above ()
3. Department
4. Section
5. Highest Qualification: Bachelor’s Degree() Masters() PhD()
Others()
6. Current Position: Court Secretary () Court clerk () court bailiff()
Magistrate grade 1() Magistrate grade 11() Senior Magistrate grade 11()
Senior Magistrate grade 1() Chief Magistrate grade 11() Chief Magistrate grade 1()
) Deputy Chief Registrar () Chief Registrar() Chief Judge()

Section B: Which ICT facilities are available in the High Court

- | | | |
|---------------------|---------|-------|
| Computers | Yes () | No() |
| Internet Facilities | Yes () | No() |
| Digital Camera | Yes () | No() |
| Scanner | Yes () | No() |
| Projector | Yes () | No() |
| Network Printers | Yes () | No() |

Section C: Which ICT facilities can you use effectively

- | | | |
|---------------------|---------|-------|
| Computers | Yes () | No() |
| Internet Facilities | Yes () | No() |
| Digital Camera | Yes () | No() |
| Scanner | Yes () | No() |
| Projector | Yes () | No() |
| Network Printers | Yes () | No() |

Section D: What factors hinder you from using ICT facilities

S/N	Questions	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
1	Lack of expertise/skills required to operate equipment					
2	No training on how to use some available equipment					
3	Lack of confidence in using the facilities					
4.	Unavailability of required ICT facilities.					
5.	Already used to the manual process.					
6.	The fear that computerizing the legal process will reduce the manual labour.					

Section E: Exploring the weaknesses of the manual processing system

S/N	Question	Always	Rarely	Not Always	Not rarely	Not at all
1	How often is a new client case registered					
2	How often are client case files updated					
3	How often are client case files removed from their storage location					
4	How often do client case files get lost					
5	How often are court summons issued					
6	How often are court warrants issued					

GENDER RECOGNITION USING LOCAL BINARY PATTERN AND NAIVE BAYES CLASSIFIER

R. S. Babatunde¹, S. O. Abdulsalam² S. R. Yusuff³ A. N. Babatunde⁴
^{1,2,3,4}*Department of Computer Science, College of Information and Communication Technology,
Kwara State University, Malete. Nigeria*

¹*ronke.babatunde@kwasu.edu.ng;*²*sulaiman.abdulsalam@kwasu.edu.ng;*
³*shakirat.yusuff@kwasu.edu.ng;*⁴*akinbowale.babatunde@kwasu.edu.ng*

ABSTRACT

Human face provides important visual information for gender perception. Ability to recognize a particular gender is very important for the purpose of differentiation. Automatic gender classification has many important applications, for example, intelligent user interface, surveillance, identity authentication, access control and human-computer interaction amongst others. Gender recognition is a fundamental task for human beings, as many social functions critically depend on the correct gender perception. Consequently, real-world applications require gender classification on real-life faces, which is much more challenging due to significant appearance variations in unconstrained scenarios. In this study, Local Binary Pattern is used to detect the occurrence of a face in a given image by reading the texture change within the regions of the image, while Naive Bayes Classifier was used for the gender classification. From the results obtained, the gender correlation was 100% and the highest accuracy of the result obtained was 99%.The system can be employed for use in scenarios where real time gender recognition is required.

Keywords: Gender, Local Binary Pattern, Naïve Bayes, Recognition

1. INTRODUCTION

Gender is a socially constructed definition of women and men. Gender can be referred to as the array of physical, biological, mental and behavioural characteristics which differentiates masculinity and femininity [1]. Automated facial gender recognition has become an interesting and challenging research problem in recent years [2].

Nowadays, researchers pay more attention to gender recognition in many potential application fields such as biometric authentication and passive demographic data collection. Gender is one of the important demographic features of human being [3]. Facial gender classification can significantly improve human identification in biometric

recognition by speeding and increasing the accuracy as it reduces the process of matching the face in the databases to nearly the half and helps in potential applications in security industry and human computer interaction [2].

Although there are diverse parts of the human face which can be used to draw out conclusion on a particular gender, according to biometric identification systems all rely upon forms of random variation among persons [11]. The more complex the randomness the better, because more dimensions of independent variation produce signatures having greater uniqueness. But while desiring maximal between-persons

variability, biometric templates also need minimal within-person variability across time and conditions. Gender recognition is one of fundamental face analysis tasks. Most of the existing studies have focused on face images acquired under controlled environment [5], [6]. However, real-world applications require gender classification on real-life faces, which is much more challenging due to significant appearance variations in unconstrained scenarios.

Gender classification is an important task which in turn can enhance the performance of a wide range of applications including identity authentication, human-computer interaction, access control, and surveillance, involving frontal facial images [7]. A large majority of gender classification approaches are based on extracting features from face images and then giving these features to a binary classifier [13].

The face of human beings tells us a lot about the identity and emotional state of the person. Recognizing the human face has been a fascinating and challenging problem, and has brought about relevant applications in many areas such as authentication for banking, security system access, and personal identification among others [8].

Local Binary Pattern (LBP) is a textural feature extractor that gives a facial representation which is independent of expression and pose artefact. The LBP operators are easy to compute hence they are suitable for real time application. LBP is robust to texture feature description and capable of representing a face in a form in which illumination variations are suppressed [9]. At a given pixel position (x_c, y_c) , LBP is defined as an ordered set of binary comparisons of pixel intensities between the center pixel and its eight surrounding pixels. The decimal form of the resulting 8-bit word (LBP code) can be expressed using Eq. 1:

$$LBP(x_c, y_c) = \sum_{n=0}^7 s(i_n, i_c) 2^n \quad (1)$$

where i_c corresponds to the grey value of the center pixel (x_c, y_c) , i_n to the grey values of the 8 surrounding pixels.

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors [10]. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, it is known to outperform even highly sophisticated classification methods (Sunil, 2015). Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$ as shown in Eq.2.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (2)$$

where $P(c|x)$ is the posterior probability of class (c, target) given predictor $(x, \text{attributes})$, $P(c)$ is the prior probability of class, $P(x|c)$ is the likelihood which is the probability of predictor given class and $P(x)$ is the prior probability of predictor.

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. A major advantage of the Naive Bayes is its efficiency in Real time Prediction. It is a fast eager learning classifier, thus, it could be used for making predictions in real time [11].

This paper aimed at developing an improved gender recognition system which employs the use of LBP to track the occurrence of a face in an image and Naive Bayes Classifier for the recognition of gender.

2. RELATED WORKS

Gender recognition systems have been in existence over the years and various researchers have brought about the use of various algorithms in order to improve gender recognition [12]. Many techniques have been used for extracting discriminative features from facial images, which can be broadly

classified into either geometric and appearance based methods [13]. [14] demonstrated that geometric feature-based methods provide similar performance to appearance-based approaches.

[4] investigated the fusion of both global and local features for gender classification. Global features were obtained using the principal component analysis (PCA) and discrete cosine transformation (DCT) approaches. A spatial local binary pattern (LBP) approach augmented with a two-dimensional DCT approach was used to find the local features. The performance of the proposed approach was investigated on Face Recognition Technology (FERET) database. The approach gives a recognition accuracy of 98.16% on FERET database. Comparisons with some of the existing techniques revealed a remarkable reduction in number of features used per image to produce results more efficiently and without loss of accuracy for gender classification.

[15] investigates gender recognition on real-life faces using the recently built database, the Labelled Faces in the Wild (LFW). Local Binary Patterns (LBP) was employed to describe faces, and Adaboost was used to select the discriminative LBP features. The research obtained the performance of 94.81% by applying Support Vector Machine (SVM) with the boosted LBP features. The public database used in this study makes future benchmark and evaluation possible.

[16] developed an automatic gender recognition algorithm based on machine learning methods. Their work consists of two stages: adaptive feature extraction and support vector machine classification. The gender recognition algorithm was based on non-linear SVM classifier with RBF kernel. To extract information from image fragment and to move to a lower dimension feature space, an adaptive feature generation algorithm which was trained by means of optimization procedure according to LDA principle was used. The results obtained

show a recognition performance of 91% on a database size of 5000 people and 79.6% on database size of 400.

[17] developed an Advanced Biometric Identification on Face, Gender and Age Recognition (ABIFGAR) algorithm for face recognition. The system yields good results on both large and small training set. It works with a training set as small as one image per person. The process consists of three phases: Preprocessing, Feature Extraction and Classification. The geometric features from a facial image were obtained based on the symmetry of human faces and the variation of gray levels, the positions of eyes, nose and mouth are located by applying the Canny edge operator. The gender and age are classified based on shape and texture information using Posteriori Class Probability and Artificial Neural Network respectively. It was observed that the face recognition is 100%, the gender and age classification is around 98% and 94% respectively.

[2] developed a facial gender recognition system using Eyes Images. In their research, feature extraction technique works only with eye and eyebrows region of the person. Three methods were used for feature extraction, which include 2D-Wavelet Transform, DCT and GLCM. SVM was used in the classification. The result from the feature extraction methods were compared and DCT gave the best recognition results over the two other methods on Face94 database used in the experiment

Although, numerous researches, including [12], [13], [15], obtained arguably high accuracy from images fetched from databases such as the Ferret database and Local Wide database, which undoubtedly would take a couple of seconds to fetch the image from the database. However, recognition systems are mostly effective in real-time scenarios such as authentication for banking, security system access, and personal identification where decisions are expected to be made in matter of nanoseconds

3. METHODS

The gender recognition system developed in this research consists of three main stages, which are:

3.1. Face tracking:

This involves determination of occurrence of face within an input set of image. Local Binary Pattern (LBP) was used for face tracking because of its robustness to texture feature description and capability to represent a face in a form in which illumination variations are suppressed [18]. The LBP was used to detect the occurrence of a face in a given image by reading the texture change within the regions of the image. Details of the operation of LBP can be found in [5]. Since the system uses a video stream, the occurrences of faces are being tracked using a green outline.

Once the LBP finds the occurrence of a face within the picture frame, a light-green boundary is created to map out the face region. From the mapped out region, the anthropometric measurements are being taken. These measurements constitute the global features used which include:

- (i) Inter-ocular distance: The distance between the midpoint of right eye and midpoint of left eye in the face image.
- (ii) Nose to Eyes: The distance between Nose tip to inter-ocular distance in the facial image.
- (iii) Lips to Nose: The distance between nose tip and the midpoint of the lips pixel in the facial image.
- (iv) Lips to Eyes: The distance between lips midpoint to inter-ocular in the facial image.

After the measurement is generated, the distance formula as used by [1] was then used to calculate the ratio of each face. Four major ratios are being calculated using Eqs. 3, 4, 5 and 6 to obtain the anthropometric measurements as itemised previously.

$$Ratio\ 1 = \frac{\text{Left eye to Right eye distance}}{\text{Eye to Nose Distance}} \quad (3)$$

$$Ratio\ 2 = \frac{\text{Eye to Nose distance}}{\text{Eye to Chin Distance}} \quad (4)$$

$$Ratio\ 3 = \frac{\text{Left eye to Right eye distance}}{\text{Eye to Chin Distance}} \quad (5)$$

$$Ratio\ 4 = \frac{\text{Eye to Lips distance}}{\text{Eye to Nose Distance}} \quad (6)$$

Once the face(s) has been tracked, there is need to bring out the discriminative features which are to be used in the gender recognition. From the face detected, the features needed which comprises of the eye region to the chin region as obtained from the four ratios in the above equations are passed on to the naïve bayes classifier. The prior probability is first calculated for both the male and the female which is then followed by the likelihood of both genders. Once this is done, each gender is calculated with respect to X which is the unknown and which forms the basis of our classification. The highest of the two becomes our classified gender.

The standard threshold values for female and male as illustrated by [1] are: ratio1>=1.1000&&ratio2>=0.7450||ratio3<=1.3714&&ratio4>=0.6404and ratio1<=1.09 &&ratio2<=0.7440 || ratio3>=1.3714 &&ratio4<=0.6400 respectively.

To obtain the prior probability, all the four ratios are considered. The prior probability of male was obtained by dividing the value obtained for male attribute based on the threshold earlier stated by the total number of ratios which is four. The prior probability for female is also calculated likewise as shown in Eq. 7 and 8.

$$Prior\ probability\ of\ Male = \frac{\text{Number of Male attributee}}{\text{Total number of Ratio}} \quad (7)$$

$$Prior\ probability\ of\ Female = \frac{\text{Number of Female attributee}}{\text{Total number of Ratio}} \quad (8)$$

Having formulated our prior probability, we obtain the likelihood. The attributes are

restricted to only Ratio 1 and Ratio 2 and these ratios are regarded as X. The likelihood is calculated using Eq. 9 and 10.

Likelihood of X given Male =

$$\frac{\text{Number of Male in X}}{\text{Total Number of Male Cases}} \quad (9)$$

Likelihood of X given Female =

$$\frac{\text{Number of Female in X}}{\text{Total Number of Female Cases}} \quad (10)$$

The final classification was carried out by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability. Therefore, after obtaining the likelihood, the product of both the likelihood and the prior probability was calculated for both gender and the highest result obtained becomes the outcome of the recognition.

3.2 Classification

This stage basically involves deciding which gender the face that has been detected falls into based on the likelihood, after the face detected has been thoroughly evaluated by the naïve bayes classifier using the Bayesian theorem.

Figure 3.1 shows the model for the gender recognition system. The face tracking was the first operation to be performed which was carried out using LBP. Next is the anthropometric feature measurement and the distance computation. The result from the distance calculation is then further passed on to the naïve bayes classifier which decides the outcome of the recognition based on the highest probability of either a male or a female gender. The gender recognition system was developed using C# Programming Language. Face images were trained in real time. The face images were then saved in a database which was built using MySQL.

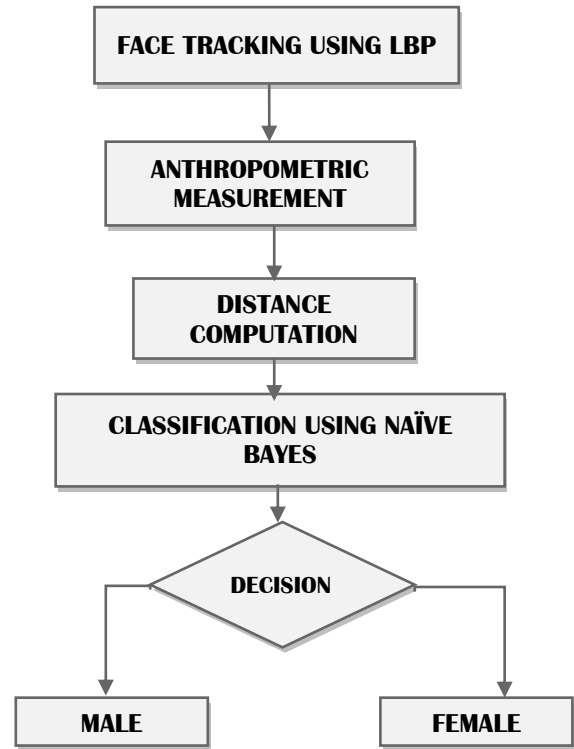


Fig. 3.1: Gender Recognition Model.

The GUI for the developed system is shown in Figure 3.2. The interface comprises of three main buttons.

- i. A start button which is used to trigger the web cam in other to get the frame of an image.
- ii. A save button which is used to save trained data
- iii. A training set button which can be used to fetch and view all saved trained images.

The interface also contains a label which is used to indicate the gender and the accuracy of the result in percentage. Once the start button is clicked from the interface, the webcam is triggered and once it is active, the detection process begins. When a face is detected from an image frame, a green box outline is drawn on the detected face and the gender in turn is immediately recognized. Then the save button will have to be used in order to save the trained face to the database.

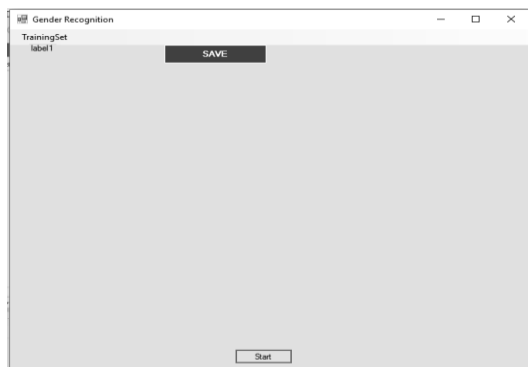


Fig. 3.2: Recognition System Interface

The application which is an executable file must run on a computer system with webcam capability which can either be an inbuilt webcam or an attached webcam and running windows operating system ranging from windows vista, Windows 7, Windows 8, Windows 8.1 or Windows 10 with Microsoft .net Framework 2.0 or 3.5. Furthermore, the application can be run on either 64-bits or a 32-bits system.

4. RESULTS AND DISCUSSION

This study was experimented on real time training of some set of facial data obtained from students of Computer Science Department of Kwara State University, Malete. The images were captured in real time via webcam whereby an individual is made to position his/her self in front of the system to acquire a frame. A green outline on the face region indicated a tracked face in the given frame. The gender recognition takes effect in matter of nanoseconds once the face is detected.

From the various results obtained, the gender of the trained data was detected accurately in terms of whether the person is a male or a female but the percentage accuracy of the gender detected is mainly dependent on the image quality in terms of the resolution and the illumination of the image and also sometimes due to variation in facial pose. The greater the illumination, the greater the accuracy obtained. Figure 4.1 shows the accurate detection of female face which was trained in a low illuminated environment.

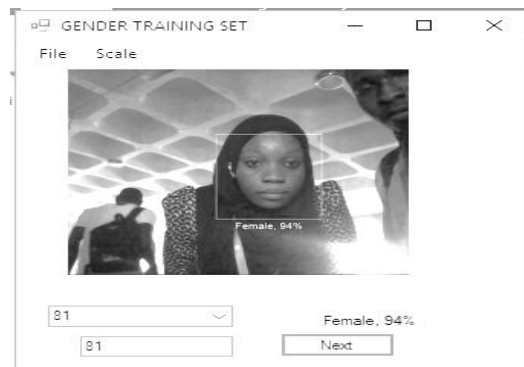


Fig. 4.1: Female training set(a)

It can be observed that the system had the right gender correlation but with an accuracy of 94% due to the low illumination as at the time the face data was trained.

Figure 4.2 shows the recognition of the same person previously shown in Figure 4.1 but with a lower accuracy, also with the correct gender correlation. It can be observed that there is slight decrease in the illumination and also the distance between the person and the camera has changed a bit. The illumination and camera distance factor only affects the accuracy of the gender recognition and not the gender correlation in particular.

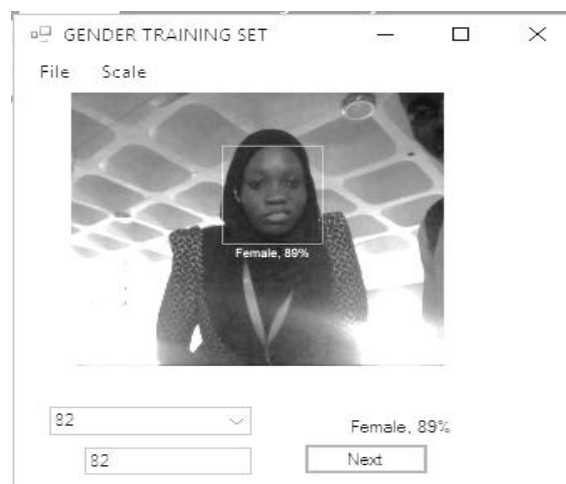


Fig. 4.2: Female training set(b)

From Figure 4.3 which captures the face of the same person as previously shown in Figure 4.1 and 4.2, it can be seen that the face has drawn closer compared to Figure 4.1 and the illumination has increased. This has caused the system to acquire a greater

accuracy.



Fig. 4.3: Female training set (c)



Fig. 4.5: Male training set (b)

Figures 4.4, 4.5, 4.6 and 4.7 shows the testing of a male student with a 99% accuracy and correct gender recognition all through the various positions of capturing in terms of distance from the camera, slight head poses, but under the same condition of illumination. It can be deduced from the result obtained in the simulation that the quality of the image in terms of resolution and distance of image from camera (i.e. distance of capture) is not a major factor when it comes to the gender correlation.

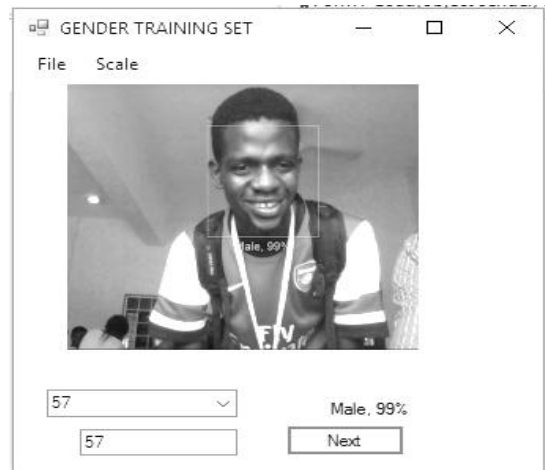


Fig. 4.6: Male training set (c)

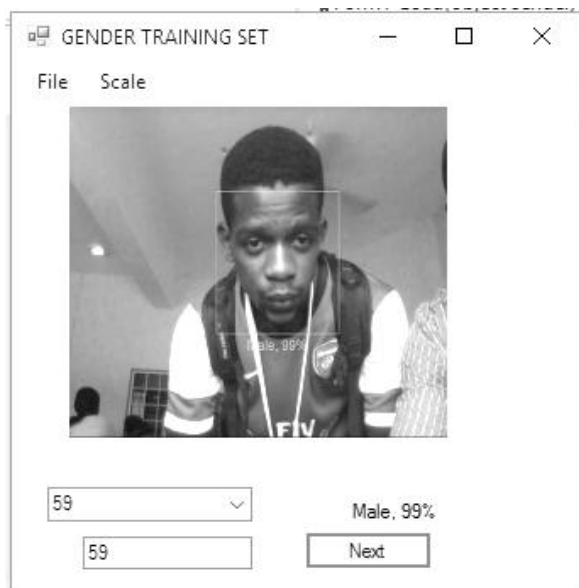


Fig. 4.4: Male training set (a)

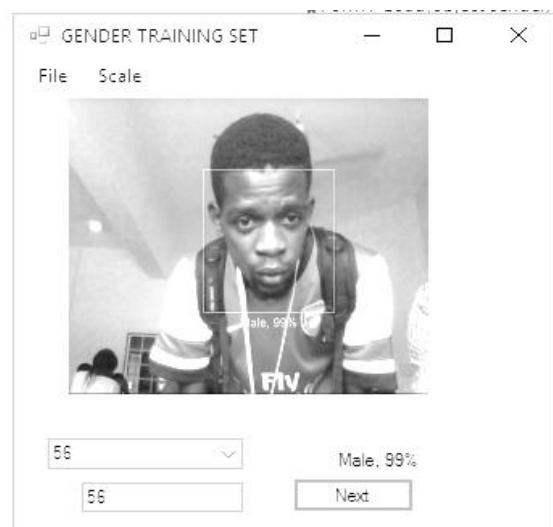


Fig. 4.7: Male training set (d)

However, different poses, illumination variation and facial expression might have a slight effect on the accuracy of the gender recognition as seen in the Figure 4.8 and 4.9.



Fig. 4.8: Male training set(e)



Fig. 4.9: Male training set(f)

5. CONCLUSION

This paper developed an automatic system capable of recognizing people's gender. From the results obtained, it can be deduced that the gender correlation is 100% and the accuracy of the result obtained is 99%, which is a great achievement being a system working in real time situation. Such systems can find useful application in different fields, such as robotics, human computer interaction, demographic data collection, video surveillance, security access control and the banking system. Further research can include the recognition of gender with high angular face pose and 3D faces.

ACKNOWLEDGEMENT

Special thanks to the Students of Kwara State University, Malete, Nigeria for their full support and cooperation during the capturing and testing of face images

6. REFERENCES

- [1] Kalam, Swathi & Guttikonda, Geetha. (2014). Gender classification using geometric facial features. *International Journal of Computer Applications*, vol. 85 No.7.
- [2] Alrashed, H. F., & Mohamed, B. A. (2013). Face gender recognition using eye images. *International Journal of Advanced Research in Computer and Communication Engineering*, 2.
- [3] Manoj, B., Pulkit, R., Annu, G. M., & Harsha, J. (2015). A Survey Paper on the Gender Recognition Techniques. Applications of Computers and Electronics for the Welfare of Rural Masses (ACEWRM). pp.22-25.
- [4] Hussain, Mirza A. Almuzaini M. M., Muhammad H., G., H Aboalsamh., and Bebis, G. (2013). Gender Recognition Using Fusion of Local and Global Facial Features. *Springer-Verlag Berlin Heidelberg. Part II, LNCS 8034*. 493–502.
- [5] Babatunde, R.S., Olabiyisi, S.O., Omidiora, E.O., Ganiyu, R.A., & Isiaka, R.M. (2015). Assessing the performance of Random Partitioning and K-Fold Cross Validation methods of evaluation of a Face Recognition System. *Advances in Image and Video Processing*, 3 (6), 19-26.
- [6] Choon, B. N., Yong, H. T., & Bok, M. G. Vision-based Human Gender Recognition: A Survey. Retrieved online at <https://arxiv.org/ftp/arxiv/papers/1204/1204.1611.pdf>
- [7] Juan, E. T., Claudio, A. P. & Kevin, W. B. (2014). Gender classification from iris images using fusion of uniform local binary patterns. *IEEE Transaction on Forensics and Security*.
- [8] Tri, H., Rui, M., & Jean-Luc, D. (2012). An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. Department of multimedia communications. Sophia Antipolis. Available at <http://dx.doi.org/10.1007/978-3-642-37410-412>
- [9] Huang, D. S., Caifeng, Ardabilian, M., Wang, Yunhong, & Chen, L. (2011). Local Binary Patterns and Its Application to Facial Image Analysis: A Survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 41 (6), 765-781.
- [10] Chai, K., Hn, H.T., Chieu, H. L. (2002). Bayesian Online Classifiers for Text Classification and Filtering. *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 97-104.
- [11] Sunil, R. (2015). 6 Easy Steps to Learn Naive Bayes Algorithm - with code in Python. Available online at <http://www.analyticsvidhya.com/blog/category/python-2/>
- [12] Lian, H., & Lu, B. (2013). Multi-view gender classification using local binary patterns and support vector machines. Department of computer science and engineering, Shanghai: Shanghai JIAO tong university.
- [13] Muhammad, H., Ihsan, U., & Hatim, A. A. (2013). Gender recognition from face images with dyadic wavelet transform and local binary pattern. *International journal on artificial intelligence tools*, 22 (6), 1-18.
- [14] Valstar, M. F., Jiang, B., Mehu M., Pantic, M., & Scherer, K. R. (2011). The First Facial Expression

- Recognition and Analysis Challenge. *IEEE Int'l. Conf. Face and Gesture Recognition(FG'11)*.
- [15] Shan, C. (2011). Learning local binary patterns for gender classification on real-world face images. *Pattern recognition letters*, 33. 1-12
- [16] Khryashchev, V., Priorov, A., Shmaglit, L., & Golubev, M. (2012). Gender Recognition via Face Area Analysis. *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*. Vol 1 pp1-5.
- [17] Jeganlal, R., Gopi, V., & Rajeswari, S. (2013). Robust automatic face, gender and age recognition using ABIFGAR algorithm. *International journal of emerging trends in Electrical and Electronics (IJETEE)*, 4(2), 41-44.
- [18] Ramchand, H., Narendra, C., & Sanjay, T. (2013). Recognition of facial expressions using local binary patterns of important facial parts. *International journal of image processing (IJIP)*, 7 (2), 1-3.
- [19] Bansal, A., Agarwal, R., & Sharma, R.K. (2014). Predicting gender using iris images. *Research Journal of recent sciences*, 3(4), 21-26.
- [20] Brian, O., & Kaushik, R. (2013). Facial recognition using modified local binary pattern and randomforest. *International journal of artificial intelligence & applications (IJAIA)*, 4 (2), 25-33.
- [21] Christopher, M. B., (2006). Pattern recognition and machine learning. *Information science and statistics, Cambridge: Microsoft Research Laboratory Ltd*.
- [22] Gupta, Varsha, & Sharma, D. (2014). A study of various face detection methods. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(5).
- [23] Nils, J. N. (2005). Introduction to machine learning. Department of Computer Science, Stanford: Stanford University. 10-12.
- [24] Rahim, A., Hossain, N., Wahid, T., & Azam, S. (2013). Face recognition using local binary patterns (LBP). *Global journal of computer science and technology graphics & vision*, 13 (4), 1-9.

**OVERVIEW OF NOSQL AND COMPARISON WITH
SQL DATABASE MANAGEMENT SYSTEMS**

V.E. Ejiofor^{1*}, K.K. Okeke²

¹*Department of Computer Science, Nnamdi Azikiwe University, Awka P.M.B 5025, Nigeria*

²*Department of Computer Science Nnamdi Azikiwe University, Awka P.M.B 5025, Nigeria*

¹*Email: ve.ejiofor@unizik.edu.ng*

²*Email: kenechu.okeke@gmail.com*

ABSTRACT

The increasing need for space in the database community has caused the revolution named NoSQL 'Not Only SQL'. With recent advancements in technology, key industries like Amazon, Google and Facebook have sought out other means to manage their resources 'effectively'. This perpetual need for robustness, cost effectiveness, flexibility, with no ambiguity (due to the volume of data involved) has moved the database industry to another phase which is the NoSQL database management system. SQL (Structured Query Language) while still holding usefulness for its ability to provide a highly consistent system still appeals to its users but both systems (SQL and NoSQL) have their various limitations which in the research and business community have caused a divide which we would call the Pro-SQL and Pro-NoSQL split. This division has made schemas/tables less and more important because NoSQL system represents a need for availability of data while SQL favours consistency. Both systems can still provide satisfaction for a user with clear understanding of what they stand for. This research has adopted an object-oriented analysis approach as the methodology and this has been used to exemplify the various techniques and systems (SQL and NoSQL).

Keywords: Availability; Consistency; Data storage; NoSQL; SQL

1. Introduction

This investigation discusses NoSQL databases in detail with the view of finding out the need for their development, their relevance or benefits which led to their creation and what informed the choice of users to opt for NoSQL databases over the relational databases and also their limitations or shortcoming. Also the research looked to discover whether NoSQL databases are replacements for the relational databases or whether they could co-exist while still maintaining the different features which make them unique for the different purposes they serve.

1.1 Overview of NoSQL Database

[1], described NoSQL as a distributed database system which does not require SQL (structured Query Language), does not have static table schemas, with no joins and can be horizontally scaled. It might also be

open-source.

According to [2], the recent migration for production purposes by some leading brands like Amazon and Google from relational to NoSQL databases is as a result of its ability to handle unstructured data such as word-processing files, e-mail, multimedia, and social media efficiently. The growing need of companies and database users for alternative options for their storage needs has moved developers to invent NoSQL databases. Different NoSQL databases use different approaches but they all have the same end purpose in common which is that they are not relational. The invention of NoSQL databases eliminates the need for the use of structured query language (SQL) which is the programming language used for querying and updating relational databases. NoSQL systems have been introduced as viable options to combat the issue of large data sets. These data sets can come in

unstructured or semi-structured forms thereby rendering the well-defined data type formation of the SQL systems impractical for the storage of these sets of data. The need for fast storage and high availability of data with subsequent faster retrieval of the stored information also aided the concept of NoSQL and big data systems development.

1.2 The Need for NoSQL Databases

According to [3], NoSQL DB is primarily arranged in uncorrelated tables. This allows for a multifaceted and fundamentally relational data model which is compacted and partitioned in various NoSQL tables. Two different partitioning policies are possible, vertical partitioning where each column of a high level relational table is stored into a separate NoSQL table and horizontal partitioning (sharding) where different sets of values are stored in different NoSQL tables. The NoSQL database allows both forms of partitioning.

NoSQL enables better and improved performance which is very important for applications with large volumes of data. Big companies like Amazon and Google have developed their own NoSQL database versions to hold their growing data and infrastructure needs. Amazon has developed the Dynamo distributed NoSQL system which uses SimpleDB as the web interface while Google has the Big Table NoSQL database and this has gone a long way in inspiring the development of new NOSQL applications. The high-performance Cassandra was developed by Facebook to help power its website while Apache created the CouchDB as an open source system which is very scalable and could be accessed from any browser. These various databases were created to meet various needs by the developers and this is as a result of the development of the web2.0 technology where relational databases cannot be used to perform multi-table join queries where massive data is involved [4]. As stated by [1], NoSQL databases offer high performance both in terms of speed, size and

high availability at the expense of losing the ACID (Atomicity, Consistency, Isolation and Durability) properties which are the major characteristics of the relational database model. This system is modelled to basically comply with features of key-value, column-orientation, document-store and graph databases. Different NoSQL databases make use of these properties based on the need of the developer and according to the purpose which the system is built to serve.

1.3 Comparison of SQL and NoSQL Databases

The traditional databases which are known as relational databases have some qualities which are attributed to them like atomicity, consistency, isolation and durability. Also they have the attribute of big feature set. According to [5], this ACID model of the relational database ensures the database is very reliable and can perform optimally with the transactions fully committing or none at all. [6] Argued that the data model in the relational databases which is normalised and its full support for the ACID properties can affect the performance of the database negatively when used for recent web activities like Facebook as joins and locks cannot perform well in a distributed environment. Joins also consume considerable resources to implement and are costly to write, debug and maintain [7]. SQL systems offer superior consistency/isolation as opposed to the NoSQL system which trades consistency for availability. This is due to real-time needs of day-to-day users who require response to answer for transactions being performed [8]. Relational databases are good for their provision of data integrity and offer of big feature set which is not an attribute of the NoSQL databases and which is due to the cost and complexity which are attached to these features [9]. Data integrity and the offer of big feature set, key attributes of relational databases, are a handicap of NoSQL databases. As argued by [6], the provision of the large feature set by the relational model can amount to needless overhead when the

query is being performed for simple tasks such as logging. According to [10], other advantages of the NoSQL database are its ability to read and write data quickly and support for mass storage, ease of expansion and low cost. In spite of these benefits, there are some inadequacies linked to NoSQL databases which are;

Lack of support for SQL (Structured Query Language): the SQL is the industry standard for querying and updating a database and the NoSQL systems do not make use of it.

Lack of transactions: transactions are the basic unit of work of a database. In the NoSQL model, there are no transactions which could be used to assess the functioning of a database.

Another setback of the traditional database is its inability to perform large amounts of read and write concurrently [1]. Being a very consistency conscious database, the relational database performs its read/write operations slowly to ensure there is no loss of information [11].

Table 1: Example of a structured database table for student records in Relational database

Student id	First Name	Last Name	Module Title	Module Code
1	kene	Okeke	Advanced Database	CSC701
2	Josh	Chukuka	Evolutionary Computation	CSC772
3	Emma Ojei	Intelligent	Agent Technology	CSC762

The student records table is fully normalised which means it is in third normal form and there is no existence of transitive dependencies within this relation. Data is held in rows in this table and there are different constraints applied to ensure integrity of this table. For example there is a primary key constraint (number (10) on the student id row which means number cannot exceed 10. Also because the student id is the primary key, the values of all the other attributes: first name, last name, module title and module code are functionally dependent on it.

Drop table student records;
 Create table student records
 (Student id number (10) not null primary key,

First name varchar2(15) not null,
 Last name varchar2(15) not null,
 Module Title varchar2(25) not null,
 Module Code varchar2(10) not null);

Commit;
Insert into student records values (1, 'Kene', 'Okeke', 'Advanced Database', 'CSC701');
 SELECT student id, first name, last name, module title, module code
 From Student records
 Where code = 'CSC772'**Example of Table creation and query in ORACLE SQL**

Table 2: Example of unstructured data for student records

Key: 1	ID: sg	First Name: Kene
--------	--------	------------------

Key: 2	Email: <u>research.compute@gmail.com</u>	Location: Awka	Age: 23
--------	--	----------------	---------

Key:3	Facebook ID: dashing	Password: brvd	Name: Emma
-------	----------------------	----------------	------------

Data for student records is unstructured. There is no defined schema holding the values together. For example in key:1 there is ID:sg and First Name: Kene while key:2 holds values for Email, location and Age. Values can be generated using the keys. The contrast between the structured and unstructured model is that the different entities like ID, First Name, location and Age would be created for both key 1 and 2 in the structured table to ensure integrity of the keys.

Syntaxes in MongoDB NoSQL

```
db.student_records.insert( { 'student_id': 1, 'first_name': Kene, 'last_name': 'Okeke', 'Module_title': 'Advanced Database', 'Module_code': 'CSC701' } );//used to insert and save data//
db.student_records.find(); //This is the select function in MongoDB database and can be used for data retrieval//
db.users.find()sort( {"module_code": 1} ); //This is an equivalent of the 'where' clause in SQL//
```

```
db.student_records.drop(); //This is used to drop a table in MongoDB//
```

1.4 Types of NoSQL Databases

The NoSQL databases are generally divided across four major categories and they are the key-value-store, the column-family, document-store and the graph databases [12]. These databases and their various attributes will be discussed in detail below.

1.4.1 Key-value Store: The concept of key-value store is based on storage of data as a pair; the key and the value [5]. The key/value is made up of two columns; one column contains the key which represents the element while the other is made up of the value which is the different parts of the element and they are data stored by a primary key [12]. As noted by [13], NoSQL databases are modelled for improved key-value stores. The structure of the key/value pair contains one or more attributes which makes it similar to the data dictionary model of the relational database. Access to the value is gained through the association of the key and the value. And the key is always uniquely identifiable in a collection [8]. This means that there is only one key.

[10] Identified the key-value model as the correspondence or mapping of a value to a key. This structure is simplified and allows faster query operations and modifications through the primary key. Read/write operations are performed faster in the key/value pair model as it allows for concurrency. [14] Noted that access to the key/value pair is purely through the primary key using methods such as PUT/GET/DELETE.

1.4.2 Example of Key-value store (Redis)

Redis is an example of the key-value store type of NoSQL database. [7], described Redis as a key-value store NoSQL system with in-built memory, which offers high availability, can perform backup and recovery and has support for ACID transactions due to its use of primary key. [10], identified several features of Redis as; periodic asynchronous caching to hard disk after the data has been committed to memory which is in-built in the Redis system and support for various operations like List and set which help in maintaining accuracy within the model. [10] Further noted that the Redis system can only hold a maximum value of up to 1GB and the main shortfall is its use of physical memory which has low capacity and therefore cannot be used for storage of big data. This lack of capacity for storage of big data reduces performance due to poor scalability of the system.

Table 3: Example of Data Representation in Key-value store

KEY	VALUE
1	Student Id: 1, Name: Kene Okeke, Year:2015/2016, Module title: Advanced Database, Module code: CSC701
2	Student Id: 2, Name: Josh Chukuka, Year:2016/2017, Module title: Evolutionary Computation, Module code: CSC772
3	Student Id: 3, Name: Emma Ojei, Year:2017/2018, Module title: Advanced Database, Module code: CSC701

The table above has all the values held in the key and the value columns. For example, Key 1 is paired with values for student id, name, year, module title and module code which are in the Value column. The key is used to insert, generate and delete the information in the value column.

1.5 Column-family: this type of NoSQL database adopts the vertical partitioning type of storage model where data are modelled to fit into columns of different NoSQL tables which make it similar to relational database model [3]. This technique allows data to be accessed through columns or column-groups. The difference between the column-store and the relational model is that tables are stores in rows for the relational database while column-oriented NoSQL databases store data in detached NoSQL tables. [3] Opined that the difference between these two models of data segregating is based on their physical infrastructural layout where data is stored in rows in the horizontally partitioned model and the column family databases where storage of data is in columns.

1.5.1 Example of Column-family (Cassandra)

As stated by [15], Cassandra is an open source database that is eventually consistent and follows the key-value store arrangement while providing high scalability across a distributed system.

Cassandra performs more write operations than it does read operations. This write operation is performed without checking the consistency state of the database and uses system versioning process of data to guarantee consistency when carrying out read operation. This leads to a reduced consistency level in the database [5].

Table 4: Example of data representation in Column-store Database

Student Id	1	2	3
Name	Kene Okeke	Josh Chukuka	Emma Ojei
Year	2015/2016	2016/2017	2017/2018
Module title	Advanced Database	Evolutionary Computation	Intelligent Agent Technology
Module code	CSC701	CSC772	CSC762

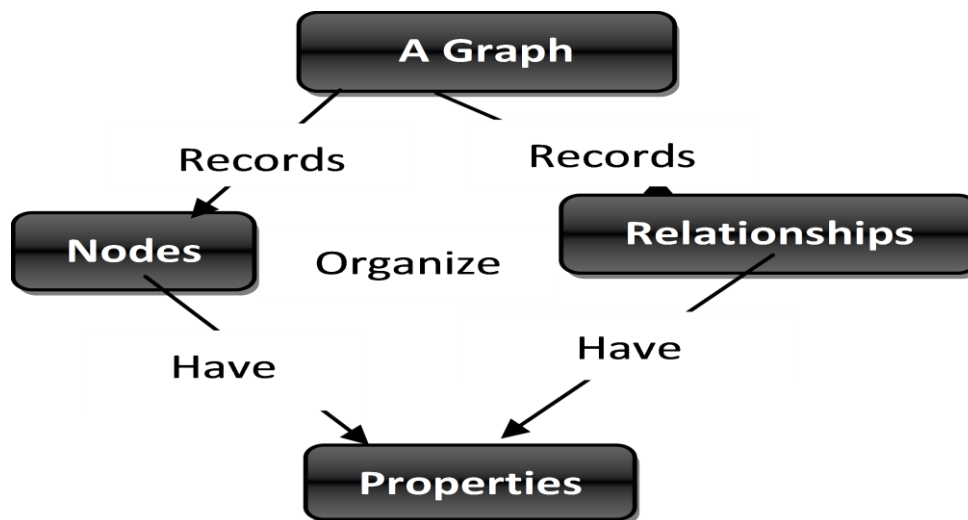
Data are stored in columns in this model. The student id row holds 1, 2, and 3 as its values. Under Name, there is Kene Okeke, Josh Chukuka and Emma Ojei in the name column, Year has 2015/2016, 2016/2017, 2017/2018, Module title is composed of Advanced Database, Evolutionary Computation and Advanced Database and Module code has CSC701, CSC772 and CSC701. The information in these columns would have been normalised and stored different in a relational database as all the information would be saved in rows.

1.6 Graph-oriented Databases

[16] Defined graph-oriented NoSQL database as the interconnection between vertices (nodes) which are linked by lines from edges. This is the joining of nodes to edges to produce an element. The nodes form the graph when they are linked with relationships. This relationship between different nodes is what produces the graph and two different nodes; the start node and end node are joined together by a relationship. [12], further defined graph database as an arrangement of nodes, edges and properties to represent and hold data. This feature enables graph database to hold complex relationships which are difficult to absorb in other databases like the many-

to-many relationships. As posited by [17], graph databases represent data in their natural format using graphical forms which show better representation rather than tabular forms. This eliminates impedance mismatch (incompatibility of database with programming language) which is the problem mostly associated with the object-relational systems which stores data in tabular form.

According to [8], the graph database faces the most challenge amongst the NoSQL databases due to its lack of support for horizontal partitioning. The horizontal partitioning which is one of the major attributes of the NoSQL databases cannot be performed in graph databases. This has been identified as a drawback in the graph-oriented NoSQL systems. The graph systems according to [7] are composed of relationships which are either static or dynamic. These relationships show connections between objects which is known as connected data for example Google and Facebook are connected data. They further claimed that graph systems have built-in access control systems used for authorisation on subsystems for applications designed for large number of end-users for/ example airlines and healthcare industries.



Example of Graph Database: Neo4j

[17], described the Neo4j as NoSQL model which is written in Java and complies with ACID properties

which means it is transaction compliant. The nodes and edges which connect to the node form the network structure and are equivalent to the entities and relationships of the relational database model. Neo4j as described by [7] is highly scalable and has persistent in-built memory which can be used for clustered systems for both single and distributed data centres.

1.7 Document-store Databases

Document-store databases are primarily developed for large data storage. It a schema-free type of NoSQL database which provides support for storage of structured and semi-structured data and these documents are stored in JSON (JavaScript Object Notation) format. The document-store NoSQL databases allow creation of convoluted data structure like arrays as an individual database entry which can be retrieved within one read procedure [14].

The document-oriented model makes every document identifiable by a unique or special key which is identified as "ID". This helps in managing its compound data structures as if they are objects linked together. The document-oriented database handles high concurrency read/write operations with ease and can access big data while simultaneously providing high availability and scalability of data. Another feature of this NoSQL system is the use of nested documents in arranging data and increasing the accuracy of stored data. According to [18], nested document views a document to be the value of another document. This makes a connection between data and is arranged in hierarchy with record as the root of a document-tree which branches out to form a schema-free database and this relationship between the record and the document-tree forms the nested document. This arrangement in a defined structure helps improve and maintain accuracy within the database.

As noted by [19], there are security issues attached to the document-store database system and this is as a result of the broad range of applications which it handles.

NoSQL systems have been identified by [20], as the database which is most suitable for Big data management because it can hold huge amounts of data in any format.

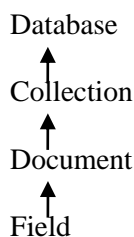


Figure 2: Layout of the Document-store Model

In the Document-store NoSQL Management System, the database holds data in a collection which is similar to tables in the RDBMS. The collection stores a list of related documents which is the equivalent of rows in the NoSQL database and the document has limitless fields which can be added to document. The field takes semblance of the column in the SQL model [5]. The difference between the relational and non-relational model is that the collection is schema-less thus eliminating the rigidity of the SQL schemas [21].

1.7.1 Example of the Document-oriented database:

According to [21], MongoDB is the most widely used document-store database and this is due to the open source nature of the system and also it is which is written in C++ programming language. MongoDB has been identified to be read-intensive which means it performs more read operations than it performs write operation.

2.0 Features of NoSQL Databases

2.1 High Availability: According to [22], availability is the capacity of the system to keep operating even when failure has been experienced. As a result of the distributed nature of the NoSQL database across many nodes, when there is failure in one node, due to data replication across these nodes, the system keeps functioning. NoSQL databases are unique for their provision of availability. It prioritises availability of data over the consistency of the database. SQL systems are high structured in nature and as a result offer a more rigid database system. The SQL systems are also available but cannot be compared to that of the NoSQL systems as a result of their semi-structured/unstructured nature. SQL systems offer consistency of data over its availability and are slower as a result of this reason.

2.1.1 Scalability: Scalability has been identified as one of the major advantages of NoSQL systems over the relational model as data could be easily spread across many servers using high-end systems and this is due to the schema-less nature of the NoSQL systems which makes the use of joins needless [23]. The join operations which are synonymous with the relational databases are very costly to perform and as they further noted, there is difficulty in performing joins on tables which are spread across a distributed system in a relational model.

2.1.2 Flexibility and Performance: NoSQL systems offer very high performance as the database can perform read/write operations very fast [5]. As they further contributed, NoSQL databases perform better in processing data across a distributed layout and can easily aggregate and update data. As argued by [13], not all NoSQL systems execute faster than the SQL databases in terms of read/write operations but due to their ability to give maximum output while holding

huge volumes of data and their flexible schema nature makes NoSQL systems unique.

2.1.3 Map-reduction:The NoSQL database allows Mapreduce to be used in querying the database directly as a form of command. As explained by [14], the NoSQL databases carry out the map-reduce function by first using themap function to process a key/value pair which divides the key/value pair into different sets of values, and then the reduce function is used to integrate the different sets of values which are related with the divided key.This map-reduce is performed using algorithm which is suitable to the type of database involved. As opined by [20], map-reduce has been a good option in processing large volumes of data as it offers symmetric execution on a sizeable number of computing connections with the reduce function performing aggregation of the

information which was delivered from the map function. This offers increased scalability and independence from the database and the programming language. SQL systems are not originally designed to perform map-reduce which came up or were introduced as a result of the large, semi-structured and unstructured nature of data. The SQL has similar operators such as FILTER and JOIN that are transformed into successions of map and reduce functions in NoSQL. Map-reduce is beyond the capability of relational database management systems and is used mainly in bigdata analytics.

- Map(k1,v1) = list(k2,v2) (1)
- Reduce (k2,list (v2)) = list(v3) (2)
- List : (a;2)(a;4)(b;4)(c;5)(b;2)(a;1) (3)
- After mapping: (a:[2,4,1]),(b:[4,2]),(c[5]) (4)
- After reducing: (a;7), (b;6), (c;5) (5)

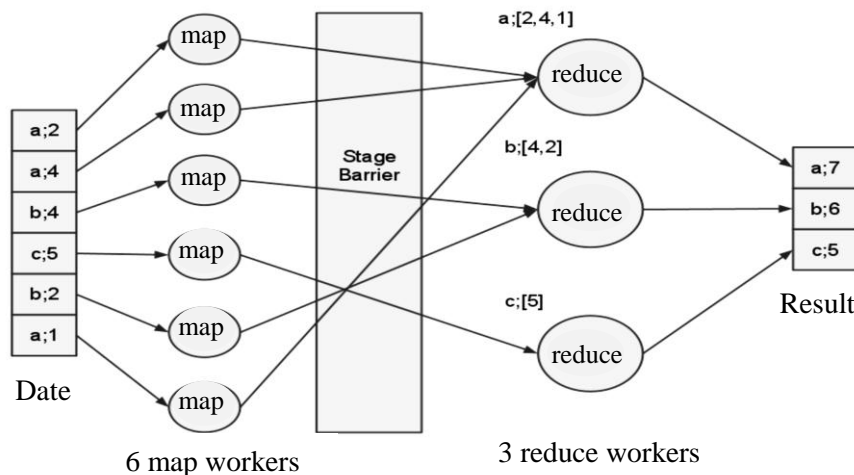


Figure 3. Equation for map reduction source [5].

2.1.4 Sharding: This is the horizontal partitioning of the database. In the relational database, partitioning is only carried out vertically which means scale-up or upward partitioning of the database. In the NoSQL database, partitioning can be in parallel or flattened (scale-out) which means more space. Sharding

increases the performance of the database as it helps balance out load when there is a rapid increase in calls of querying to be performed by the database which depends in part on configuration and to a large extent on horizontal partitioning of the system [7].

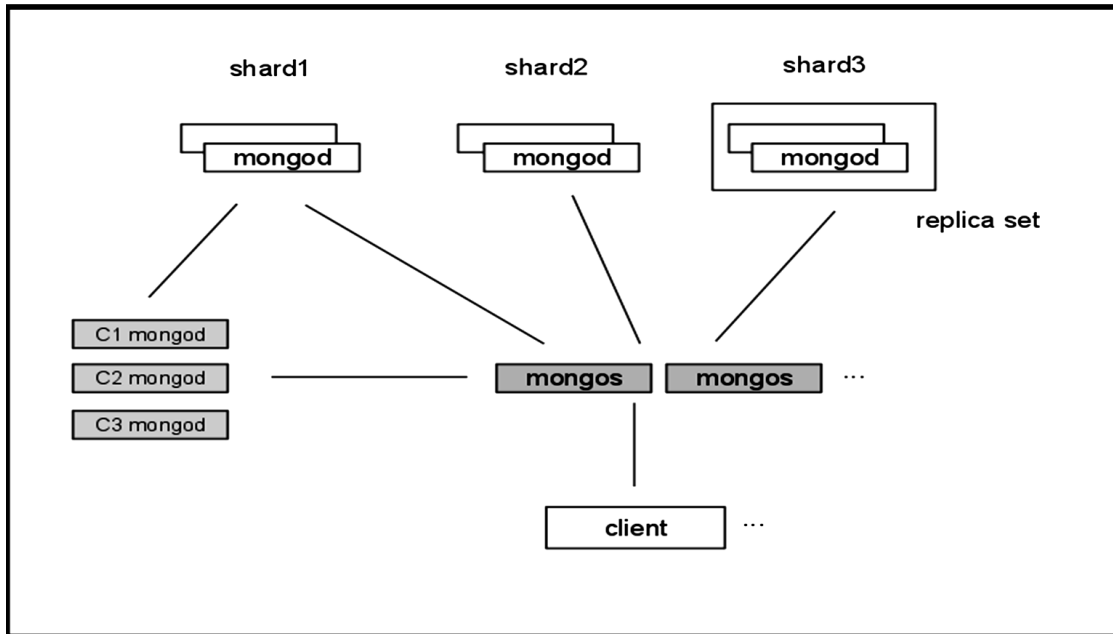


Figure 4: Example of sharding performed in mongodb source: [5].

2.1.5 Replication:[6], acknowledged that NoSQL databases offer automatic data replication which means that in case of any temporary failure across a node in the database, the entire system will maintain availability and effective load balancing with replacement of that node by replica servers while the database performs recovery of the failed node. As described by [5], some NoSQL systems like Cassandra use a duplication system known as Multi-Version Concurrency Control (MVCC) which stores replicated versions of the same data and matches the various versions before merging them. The MVCC is most useful when applied to a distributed system as it avoids locks and handles concurrent write operations efficiently by providing eventual consistency. [5]Further noted that there is an inconvenience attributable to MVCC model of replication as it has to delete old entries from time to time which lead to loss of time in the overall database throughput. Another form of replication is the Master/Slave where one server which is considered the master performs read/write operations and another server (slave) replicates data and handles read and backup operations. This form of replication is used mainly for NoSQL systems like Mongo DB which uses locks on the database [5].As some document-store NoSQL databases like MongoDB perform more read operations, [4], stated that they make use of replica sets which allows for a greater number of slaves while only one server which is write intensive carries out write operations for the management system. Replication is one of the good features available in SQL Server and this is termed transactional replication. Transactional Replication is used when DML or DDL schema changes implemented on an object of a database on one server requires to be reproduced on the database residing on another

server. This process offers high throughput and occurs almost in real time (i.e. within seconds).

2.1.6 Low Latency: [21] noted that the NoSQL model is designed to conform to low latency rate within the management system. According to [8], due to the replication of data in NoSQL, a failure goes unnoticed as replica servers generate the exact information on the failed node. The NoSQL system considers when there is partition, availability and consistency of the database and when the system performs normally without partition, it considers latency and consistency.

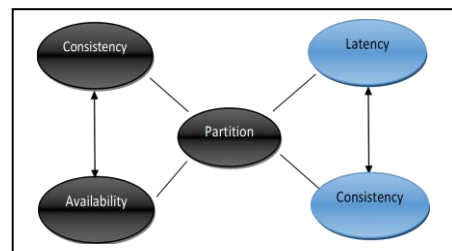


Figure5: Example of PACELC model source: [8]

As inferred by [8], the PACELC (Partition, Availability and Consistency, Else, Latency and Consistency) theorem inculcates all the characteristics of the NoSQL management system. These features are interwoven in their functionality and are meant to arrive finally at a highly available database. Mapreduce produces scalability for the database, scalability allows for improved performance as data is processed with ease and speed. Sharding (horizontal partitioning) allows for replication where data is reproduced and spread across many servers which is based mainly on the schema-free (flexible) nature of the NoSQL systems

and all these measures are put in place to ensure reasonably consistent and maximally available provision of a database that is considered to be

Table 5: Feature of NOSQL

Features	Redis	Cassandra	MongoDB	CouchDB	RavenDB	Neo4j
Language Written in:	C	Java	C++	Erlang	C#	Java
Storage Capacity	Physical storage in Disk (small storage capacity)	Memory (large capacity)	Memory (large capacity)	Memory (large capacity)	Memory (large capacity)	Memory (large capacity)

Data model	Key-value	Column	Document	Document	Document	Graph
Sharding (horizontal Partitioning)	Not supported	Performs sharding	Performs sharding	Not supported	Performs sharding	Not supported

License type (source)	BSD	Apache	AGPL	Apache	AGPL	AGPL
	Open source	Open source	Open source	Open source	Open source Commercial	GPL Open source

Fast Concurrent read/write	Fast concurrent read/write	Fast concurrent read/write	Fast concurrent read/write	Fast concurrent read/write	Fast concurrent read/write	Fast concurrent read/write
	Read-intensive	Write-intensive	Read/write intensive	Read-intensive	Read-intensive	Read-intensive
CAP Theorem (Transaction)	Availability, Partition-Tolerance	Availability, Partition-Tolerance	Availability, Partition-Tolerance	Availability, Partition-Tolerance	ACID	Consistency, Availability
Consistency	Model Strong consistency	Eventual consistency	Strong Consistency	Eventual consistency	Strong Consistency	Strong Consistency
Indexes (secondary indexes)	Global primary indexes(no secondary indexes)	No indexes	Has indexes	Has indexes	Has indexes	No indexes
Ad-hoc Query	None	HIVE, PIG	BSON based format (MongoSQL)	Lucene, Cloudbant	Limited, built-in (JSON format)	Cypher

The features above are summary of the various characteristics of the different NoSQL databases which have been considered for use by the developer. For querying in the NoSQL systems outlined, various APIs are used to achieve this purpose as there is no unified query language for NoSQL systems. The relational model has a unified query Language which is SQL and also allows for query optimisation.

3.0 Limitations of NoSQL

3.1 Heterogeneous data structure

Unrelated data such as emails, multimedia and blogs are stored in a schema-less table which can come from different sources for example the social media and this means varied structures of data are held together in one table. As the table grows or expands, it becomes an issue for the database to fully provide a uniform application interface which can encompass these diverse data combination [12].

The heterogeneity is as a result of different data models which are used by different NoSQL data-stores. Even when the data model is the same, there are variations due to different implementations. There is also the issue of the differences in query language used and the type of consistency model used by a particular NoSQL Database Management System. Different NoSQL models apply different CAP formation and the combination by one NoSQL management system might not be supported by another model [24].

3.1.1 Near inconsistency of the database

As noted by [10], Professor Eric Brewer in 2000 proposed the CAP theorem which stands for (Consistency, Availability and Partition tolerance). According to [5], the CAP model grants that in a joint-data scheme, only two out of the three features can be satisfied at a particular point in time within the database. As they further explained, there are three possible configurations which are; consistency and partition tolerance, availability and partition tolerance and the last consistency and availability which is very difficult to combine.

In NoSQL system, availability and partition tolerance are ranked higher and valued over consistency. The system is satisfied with eventual consistency [15].

Also the BASE (Basically Available, Soft state, eventually consistent) pattern of the NoSQL databases proves that NoSQL systems are more disposed towards availability of data than they are towards consistency of the data involved [7].

3.1.2 Varying Query Model: Unlike the relational model which uses the SQL (Structured Query Language) as its unified language for querying RDBMS (Relational Database Management Systems), different NoSQL vendors have different languages which they use to access their database. There is no standard interface for the NoSQL database management systems [24]. According to [25] this variability in data format leaves an overhead

as a result of different API which would be used in analysis of data.

3.1.3 Security issues in design model of NoSQL systems

According to [26], NoSQL systems with their inherent auto-sharding for reliable high performance and good load balancing were not ab initio designed with security as one of the key features. Sharding as they further noted pose security risks as unencrypted data is replicated and distributed across many servers in different locations allowing for vulnerability of data as unauthorised users can gain access to information. This breach could also be as a result of communications within a network that is not very secure.

[19], stated that amongst the features of the NoSQL systems is the lack of referential integrity which is maintaining integrity constraints for the foreign key and also very little support for security at the database level. This lack of support for security within the management system means there could be breaches to data which has been stored in the database.

Amongst the incumbent problems of the NoSQL databases is the prevalence of Denial of service problem [19]. Attackers gain access to IP addresses of users by sniffing the network and divert resources to pseudo connections thereby denying legitimate users the service allotted to them.

3.2 Security enhancement for NoSQL database:AS

proposed by [26], some techniques also identified as loopholes such as authentication, access controls, secure configuration, data encryption and auditing can be improved upon for a secure NoSQL database system.

3.2.1 Authentication: Authentication is the verification of identify of users to ensure that the rightful users are granted access to database resources. Authentication can be for a single user or group access or verification between servers. To secure the sharded NoSQL database using authentication [26] recommended some authentication methods like use of protocols such as SSL (Secure Socket Layer) and SSH (Secure Shell) which are efficient cryptographic models. Also, certificate based authentication where every certificate is verified and Password based authentication methods could be used.

3.2.2 Access Controls: This is a process for ensuring that only authorised users gain access to database resources. Some proposed access control techniques by [26] are RBAC (Role Based Access Control), DAC (Discretionary Access Control) and also MAC (Mandatory Access Control). These various models listed can be applied based system configurations to ensure restricted access only for functions specified for a particular user.

3.2.3 Secure Configuration: Faulty configurations at the Operating System, in the database or the application layer can lead to breach of security through the entire database. The suggested configuration runs from backup to updates and services. Proper configuration of ports, files and directories and protocols was also recommended.

3.2.4 Encryption of data: Encryption of data is used to provide secrecy of information within the database (data-at-rest) and across the network (data-in-transit) using mathematical algorithms. Some cryptographic standards proposed are DES (Data Encryption Standard), AES (Advanced Encryption Standard) and could be used to protect data within the database. Some techniques for securing data which is sent across the network are IPsec, SSL, TLS and SSH.

3.2.5 Auditing: Audit trails which can be used to monitor activities performed in the database can be used to enhance security of data in the database. Auditing stands for monitoring and noting activities carried out database users. It can be used to detect infiltration attempts by attackers. This is period check on connections and activities.

4.0 Conclusion and Future Work

The investigation carried out shows that NoSQL systems are not replacements for the relational database systems as each could be used to perform specific purposes with optimum output. For some companies with need for a very consistent system

which can hold well-structured data sets, the relational databases are still the best option for them. The NoSQL systems with their speed, availability and fail-over options serve the purpose of delivering run-time purposes of good throughput.

The choice of database system would be based entirely on users' needs and discretion. For example, a user in need of a database to store information about staff and their salaries might go for a SQL system which can take care of the needs as required. Alternatively a blogger requiring a database will be best suited to choose NoSQL as it would be the option which can provide satisfactory dispensation of tasks.

“As NoSQL declared, they will provide more options for different situations which are suited for their applications, so it is “Not only SQL” [15].

As comparison was made between SQL and NOSQL systems, a comprehensive work on how the system could be combined is proposed.

The peculiarities of the SQL system which is based strongly on consistency made possible by the ACID (Atomicity, consistency, isolation and durability) properties could be combined with the BASE (Basically available, soft state eventually consistent), PACELC (Partition, Availability, consistency, Else, Latency and consistency) and CAP (Consistency, Availability and partition-tolerance) theorems of the NOSQL system which offers availability as its priority. This combination will leverage an interoperable system.

REFERENCES

- 1) B. Tudorica and C. Bucur (2011). “A comparison between several NoSQL databases with comments and notes.” *RoeduNet International Conference (RoEduNet) 2011 10th*, Pp.1-5.
- 2) S. Ramanathan, S. Goel and S. Alagumalai (2011). “Comparison of Cloud database: Amazon's SimpleDB and Google's Bigtable.” *International Conference on Recent Trends in Information Systems (ReTIS) 2011* pp.165—168
- 3) S. Lombardo, E. Di Nitto, D. Ardagna (2012). “Issues in Handling Complex Data Structures with NoSQL databases.” *IEEE 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. Pp. 443-448.
- 4) Z. Wei-ping, L. Ming-Xin and C. Huan (2011). “Using MongoDB to implement textbook management system instead of MySQL.” *IEEE 3rd International Conference on Communication Software and Networks (ICCSN), 2011*. Pp.303--305.
- 5) L. Bonnet, A. Laurent, M. Sala, B. Laurent, and N. Sicard (2011). “Reduce, you say: What nosql can do for data aggregation and bi in large repositories.” *IEEE 22nd International Workshop on Database and Expert Systems Applications (DEXA)*. Pp.483-488.
- 6) R. Hecht and S. Jablonski (2011). “NoSQL Evaluation: A use case oriented survey.” *Proceedings of International Conference on Cloud and Service Computing (CSC)*. Pp.336-341.
- 7) V. N. Gudivada, D. Rao, and V.V. Raghavan (2014). “NoSQL Systems for Big Data Management.” *IEEE World Congress on Services (SERVICES)*. Pp. 190-197.
- 8) M. Indrawan-Santiago (2012). “Database research: Are we at a crossroad? Reflection on NoSQL.” *IEEE 15th International Conference on Network-Based Information Systems*, Pp.45-51.
- 9) R. Pettersen, S. Valvag, A. Kvalnes and D. Johansen (2014). “Jovaku: Globally Distributed Caching for Cloud Database Services Using DNS.” *IEEE 2nd International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), 2014*. Pp.127--135.
- 10) J. Han, E. Haihong, G. Le and J. Du (2011). “Survey on NoSQL database.” *IEEE 6th International Conference on Pervasive Computing and Applications (ICPCA)*. Pp.363-366.
- 11) P. Xiang, R. Hou and Z. Zhou (2010). “Cache and consistency in Nosql.” *IEEE International Conference on Computer Science and*

Overview of NOSQL and Comparison with SQL Database Management Systems
V.E. Ejiofor and K.K. Okeke

- Information Technology (ICCSIT)*. 6, Pp.117-120.
- 12) D. Jayathilake, C. Sooriaarachchi, T. Gunawardena, B. Kulasuriya and T. Dayaratne (2012). "A Study Into Capabilities of NoSQL Databases in Handling a Highly Heterogeneous Tree." *IEEE 6th International Conference on Information and Automation for Sustainability (ICIAfS)*. Pp. 106-111.
 - 13) Y. Li and S. Manoharan (2013). "A performance comparison of SQL and NoSQL databases." *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. Pp.15-19.
 - 14) S.S. Nyati, S. Pawar and R. Ingle(2013). "Performance evaluation of unstructured NoSQL data over distributed framework." *IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Pp. 1623-1627.
 - 15) G. Wang and J. Tang (2012). "The NoSQL Principles and Basic Application of Cassandra Model." *IEEE International Conference on Computer Science & Service System (CSSS)*. Pp.1332-1335.
 - 16) A. Castellort and A. Laurent (2013). "Representing history in graph-oriented NoSQL databases: A versioning system." *IEEE 8th International Conference on Digital Information Management (ICDIM)*. Pp.228-234.
 - 17) K. Kaur and R. Rani (2013). "Modeling and querying Data in NoSQL Databases." *IEEE International Conference on Big Data*. Pp.1-7.
 - 18) Z. Jiang, Y. Zheng and Y. Shi (2013). "Document-Oriented Database-Based Privacy Data Protection Architecture." *IEEE 10th International Conference on Web Information System and Application (WISA)*. Pp. 19-22.
 - 19) L. Okman, N. Gal-Oz, Y. Gonen, E. Gudes and J. Abramov (2011). "Security issues in Nosql databases." *IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. Pp.541-547.
 - 20) K. Grolinger, M. Hayes, W.A. Higashino, L'Heureux, D.S. Allison, M.A.M. Capretz (2014). "Challenges for MapReduce in Big Data." *IEEE World Congress on Services (SERVICES)*. Pp.182-189.
 - 21) A. Boicea, F. Radulescu, and L. Agapin (2012). "MongoDB vs Oracle - database comparison." *IEEE 3rd International Conference on Emerging Intelligent Data and Web Technologies*. pp.330—335.
 - 22) S. Benefico, E. Gjeci, R.G. Gomasasca, E. Lever, S. Lombardo, D. Ardagna, and E. Di Nitto (2012). "Evaluation of the CAP Properties on Amazon SimpleDB and Windows Azure Table Storage." *IEEE 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. pp. 430-435.
 - 23) N. Leavitt (2010). "Will NoSQL Databases Live Up to Their Promise?" *IEEE Computing*, 43(2), Pp.12-14.
 - 24) H.M.L. Dharmasiri and M.D.J.S. Goonetillake (2013). "A federated approach on heterogeneous NoSQL data stores." *IEEE International Conference on Advances in ICT for Emerging Regions (ICTer)*. pp. 234-239.
 - 25) R.K Lomotey and R. Deters (2014). "Towards Knowledge Discovery in Big data." *IEEE 8th International Symposium on Service Oriented System Engineering (SOSE)*. Pp. 181-191.
 - 26) A. Zahid, R. Masood and M.A. Shibli (2014). "Security of shaded NoSQL databases: A comparative analysis." *IEEE Conference on Information Assurance and Cyber Security (CIACS)*. Pp. 1-8.

ADAPTIVE GUARD CHANNEL ALLOCATION SCHEME WITH BUFFER FOR MOBILE NETWORK

¹Ojesanmi O. A., ¹Oyedele O., ¹Vincent O.R. and ²Olayiwola M.

¹Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria.

²Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria.

*ojesanmioa@funaab.edu.ng, sanyaoyedele@yahoo.com, vincent.rebecca@gmail.com.
olayiwolam@funaab.edu.ng*

ABSTRACT

The devastating effect congestion has on the quality of service delivery and overall network performance demands an utmost attention. This certainly calls for taking some expedient measures to deal with congestion so as to salvage the network from total collapse. In this paper, an adaptive guard channel allocation scheme with buffer to handle resource assignment in mobile network is presented. The scheme uses a dynamic reservation system that adapts to network characteristics for efficient allocation. The available channels are divided into two: open channel and reserved channel. The open channels are used by both new and handoff calls when channels are available while only handoff calls are allowed to use the guard channel when there are no idle channels at the open. The input traffic rate determines the threshold of the guard channel. A simulation program written in Java programming language evaluates the performance of the scheme based on blocking/dropping probabilities of both calls. Results of the evaluation is described using descriptive statistics such as bar charts. The proposed scheme would reduce congestion and improve quality of service delivery in mobile network.

Keywords: Guard Channel, Quality of Service, Handoff Call, New Call, Buffer, Blocking Probability

1.0 INTRODUCTION

There is undoubtedly a high demand for wireless connectivity at the moment. In spite of that, more users have to be given access to the limited available bandwidth for more revenue to be generated. This can only be guaranteed when the bandwidth available are properly managed, because improper management of it will drastically reduce the revenue being generated. Therefore, proper allocation of bandwidth and efficient utilization of same have become issues of paramount importance. [16][28][30]. The major problems of the mobile wireless network as customers become many are network congestion and signal quality degradation. These issues ceaselessly crave for more researches to bring about improvement in network performance. Many attempts have been made to proffer lasting solution to congestion. These attempts are either to avoid congestion or manage it. The congestion control

methods in Global System for Mobile communication (GSM) include the following: token bank, automatic call gapping and Call Admission Control (CAC). CAC is adjudged as the best and that is why it is considered for this research work [6][8].

In call admission control, the principle is to guarantee the quality of service (QoS) of every connection to the network by efficiently managing the available network resources. An efficient call admission control technique should exhibit the following characteristics: reliable priority assigning strategy for calls of different service classes; relatively low call blocking probability; efficient and fair allocation of network resources; increase in network throughput and congestion prevention. The admission decision, at times, is being controlled by the QoS requirements of the network users and not only on the network resources available [11]. For call admission control to

be effective, there must be an efficient channel allocation scheme. Many channel allocation techniques have been proposed [9][17], but they did not make provision for buffer which can bring about significant reduction in blocking of handoff calls and thereby minimize congestion.

In this research, a guard channel allocation with buffering of handoff calls is employed. Whenever there is no available channel for arriving handoff call, the call, instead of being blocked is buffered and later allocated channel when there is free channel. This greatly reduces the handoff dropping probability and it is a significant improvement on other conventional static strategies which automatically allows a handoff call to be blocked when there is no available channel.

2.0 RELATED SCHEMES

Several channel allocation schemes have been studied. In Adaptive Channel Allocation Scheme, handoff calls are dynamically allocated channels in consideration of certain past period in the network. An important factor in getting good Quality of Service hinges on the selection of number of guard channels exclusively reserved for handoff calls. Specific number of guard channels are needed to be allocated for different type of traffic load and mobility factor. As the traffic load changes with time, it is essential that the number of guard channels also changes. The Adaptive Channel Allocation Scheme is designed to search for the maximum number of guard channels to be exclusively reserved at each base station for handoff calls. The scheme ensures optimal utilization of the available network resources, and load balancing in the network traffic [3] [15] [19].

In dynamic channel allocation scheme, channels are not pre-assigned to the cells within the cellular network. The available channels are stored in a central pool which are shared among the requesting calls in each cell. A co-channel reuse constraint must be satisfied before a channel is occupied by a call in any cell. By design, all cells within the same region gain access to the channels kept in the pool. The request call is assigned channel by an algorithm that takes into consideration some important factors, namely: the possibility of future blocking within the cell, the reuse distance of the channel and other cost functions. A call is assigned channel, and at the expiration of that call, the channel is sent back to the pool, or re-assigned to the same cell site that was previously in control of the channel. This strategy considerably reduces the call blocking probability and greatly enhances the performance of the system due to increase trunking capacity. The key achievements are real time data, traffic distribution and radio signal strength (RSSI) and proper utilization of channel [4] [16] [21] [24].

Channel borrowing scheme ensures that a cell (an acceptor) that has exhausted all its assigned channels can freely borrow channels from the nearby cells (donors) to accommodate or allow handoffs. This is possible in as much as the borrowing does not disrupt

the existing calls. Other cells are prevented from using an already borrowed channel, and this is referred to as channel locking which rubs positively on the performance of channel borrowing schemes. It is possible to borrow from an adjacent cell with the largest number of free channels or alternatively pick the first free channel available for borrowing through the use of a search algorithm [2] [7] [19].

Guard channel prioritization scheme significantly reduces the call dropping probability by reserving in each cell a reasonable number of channels for handoff calls. The handoff and new calls then share equally the remaining channels. The moment the number of free channels is equal to or less than the predefined threshold, guard channels are established. With this, the new calls are not served, but only handoff calls are assigned channels until all channels are filled up [2] [8].

In a hybrid scheme, both fixed and dynamic channel allocation schemes are combined. With the fixed allocation strategy, each cell is assigned a fixed set of frequency channels. A call is served only when there are unoccupied channels within the cell where a call request is sent. Also, with dynamic allocation, channels can be dynamically borrowed from other cells [12] [23].

Dynamic load balancing technique with CAC ensures that the Call Admission Control technique (CAC) is combined with load balancing strategy. CAC determines the condition for accepting or rejecting a call based on the available network resources. Load balancing involves dividing network traffics between network interfaces. This seeks to efficiently utilize available resource, improve throughput and prevent overloading. With this, a good quality of service is achieved in the presence of large volume of traffic, and congestion is prevented [6].

3.0 METHODOLOGY

3.1 Model Assumptions

The assumptions about the proposed architecture are explained. The originating and handoff calls' arrival pattern follows the Poisson process with mean rate λ_{ON} for originating calls and mean rate λ_{HN} for handoff calls. The service time or departure rate has an exponential distribution with mean rate μ . The total available channels in the cell is A , and it is categorized into open channel, whose boundary is A_c , and guard channel whose boundary is A . There is a single queue for handoff calls and its capacity is w . At the completion of a call or when the user moves away from the cell, the queue is cleared.

3.2 System Architecture

The total available channels are divided into two: the guard channels, which are exclusively reserved for handoff calls; and the open channels, which are for both the new calls and handoff calls. The admission of new calls occurs if there are free open channels. If not, they are blocked. On the other hand, handoff calls are

admitted if there are either free guard channels or open channels. If there is neither, handoff calls which can either be real time or non-real time are buffered. The process of queuing or buffering handoff calls can be summarized this way. Real time and non-real time handoff calls are buffered if there are no channel available in the destination cell, and are queued pending the time the channel is available. In a situation when there is high demand for real time and non-real time handoff, calls are denied to stay on queue because of the limited buffer size. The system architecture and system flowchart are shown in Figures 1 and 2 respectively. As some of the new and handoff calls are

leaving the channel, the handoff calls in the buffer start occupying those channels. Real time handoff calls are given higher priority than the non-real time handoff calls in the allocation of the available channels. But as each of the real time handoff calls is being served, the probability of the non-real time handoff calls receiving service also increases (when two real time calls are served, then one non-real time call is also served). This is to prevent the non-real time handoff calls from waiting endlessly in the buffer. This greatly reduces the dropping probability of the handoff calls. The system procedure is given in Algorithm 1.

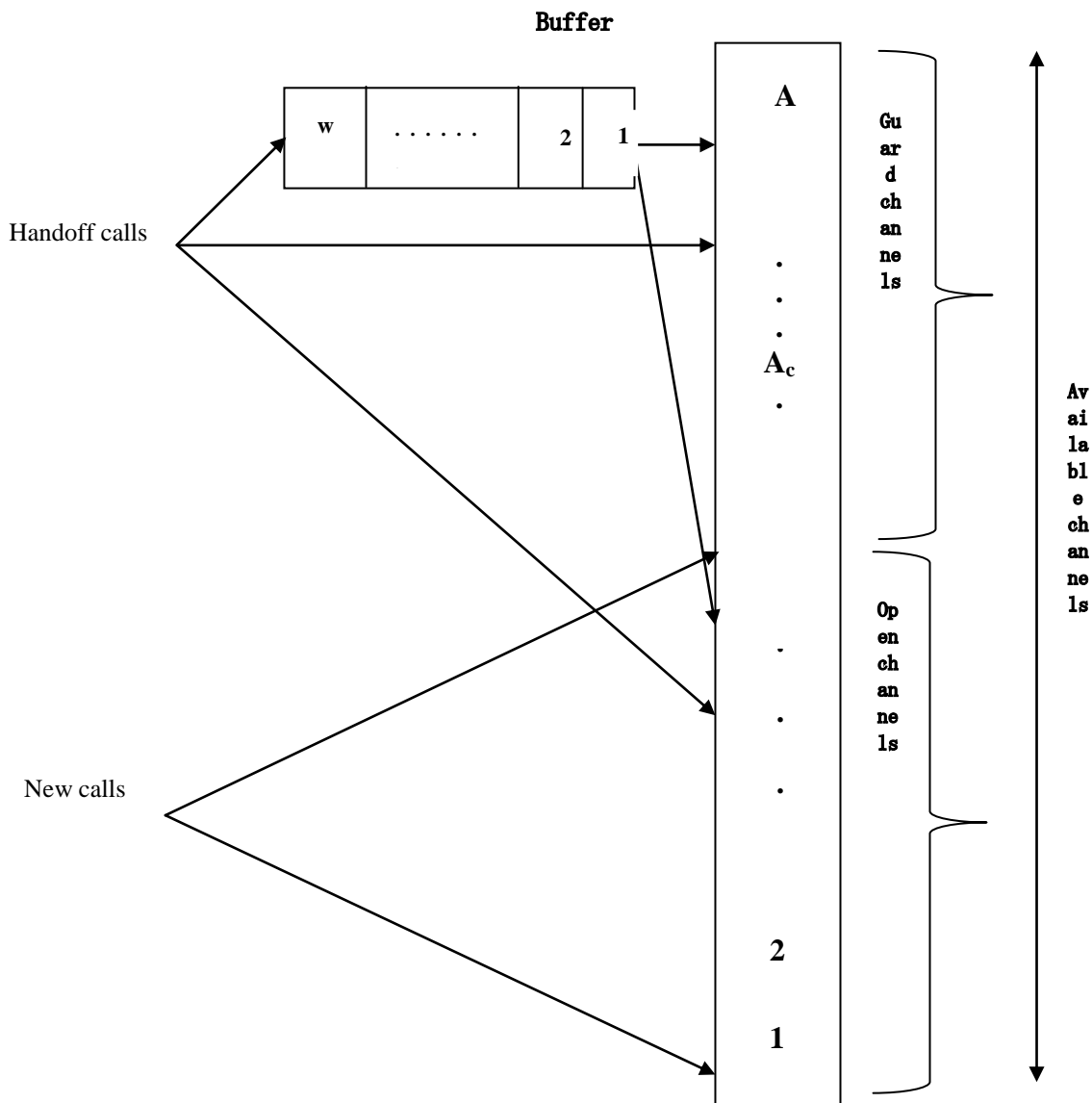


Figure 1: System Architecture

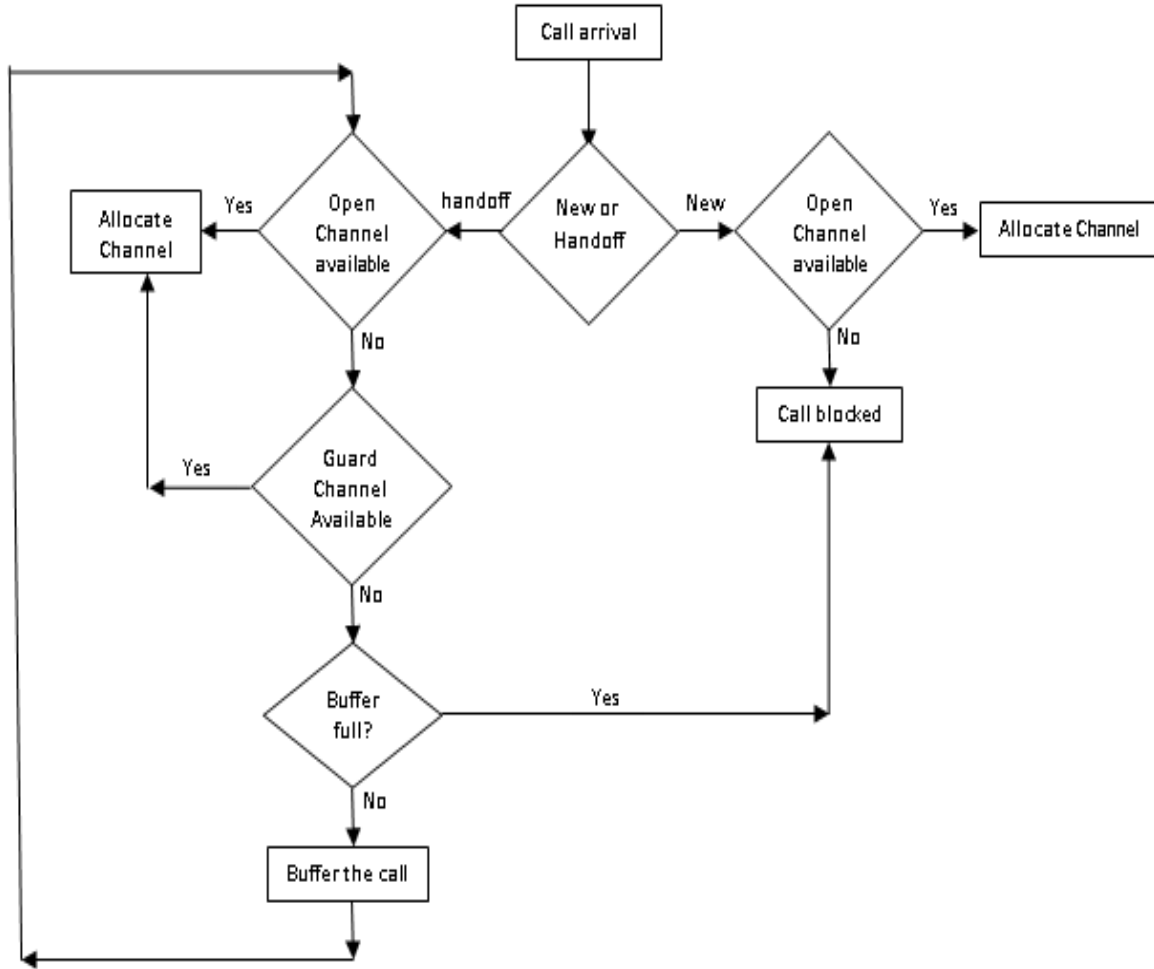


Figure 2: Flowchart for the proposed model

Algorithm 1: Channel Assignment Algorithm

INPUT: Call request (New calls, Handoff calls)
OUTPUT: (New call probabilities, Handoff call probabilities)
 1: if (incoming request is new call or handoff call)
 2: if (there is a free channel in the open) then
 3: allocate the free channel
 4: else
 5: if (handoff call)
 6: if (free channel in the guard)
 7: then allocate free channel
 8: endif
 9: endif
 10: if (no channel is available)
 11: is buffer full?
 12: If No then
 13: put handoff call in buffer
 14: else
 15: block the handoff call
 16: endif
 17: If (there is free channel again)
 18: allocate the free channel to handoff call
 19: endif
 20: endif
 21: endif

3.3 MATHEMATICAL MODEL

The system in consideration consists of many cells. The number of channels within each cell is A . The time it takes a call to expire after being assigned a channel is called channel holding time and its mean rate is μ , having also an exponential distribution. The arrival patterns of originating and handoff calls follow Poisson processes, and their mean rates are λ_{ON} and λ_{HN} respectively. It is worthy of note that all the cells within the system are of the same properties, but concentration is on a single cell. Calls that are just being initiated are referred to as originating calls or new calls. When a mobile station gets closer to a cell from a neighboring cell with a very strong signal, a handoff call is produced. Handoff calls are usually prioritized over new calls. As a result of that, some channels, specifically A_R channels in this case, are exclusively assigned out of the available A channels. The remaining channels, $A_C = (A - A_R)$ are co-shared by both the originating and handoff calls. When the number of available channels is less than or equal to A_R , a new call is certainly blocked. When there is unavailability of channel and the buffer is full, a handoff call is blocked.

The state q ($q = 0, 1, \dots, A$) of a cell can be defined as the number of ongoing calls for the base station of that cell. If $P(q)$ denotes the steady-state

probability that the base station is in state q , then probabilities $P(q)$ can be derived for birth-death processes. Figure 4 illustrates the resulting state transition diagram.

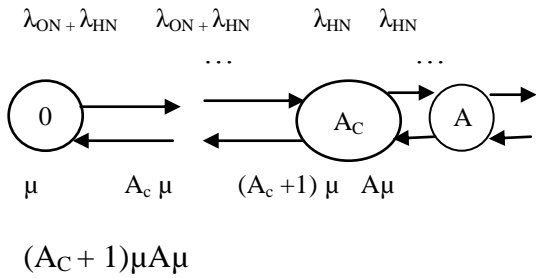


Figure 4: State transition diagram for the system model

The state balance equations derived from figure 4 are:

$$\begin{cases} q\mu P(q) = (\lambda_{ON} + \lambda_{HN}) P(q-1) & 0 \leq q \leq A_C \\ q\mu P(q) = \lambda_{HN} P(q-1) & A_C < q \leq A \end{cases} \quad (1)$$

where λ_{ON} and λ_{HN} represent the arrival rate of originating and handoff calls respectively

$$\sum_{q=0}^A P(q) = 1 \quad (2)$$

The steady-state probability $P(q)$ can be deduced as follows:

$$P(q) = \begin{cases} \frac{(\lambda_{ON} + \lambda_{HN})^q}{q! \mu^q} P(0) & 0 \leq q \leq A_C \\ \frac{(\lambda_{ON} + \lambda_{HN})^{A_C} \lambda_{HN}^{q-A_C}}{q! \mu^q} P(0) & A_C \leq q \leq A \end{cases} \quad (3)$$

From Figure 4, it is observed that, steady state probability that the system is in state ‘0’

$$P(0) = \sum_{q=0}^{A_C} \frac{(\lambda_{ON} + \lambda_{HN})^q}{q! \mu^q} + \sum_{q=A_C+1}^A \frac{(\lambda_{ON} + \lambda_{HN})^{A_C} \lambda_{HN}^{q-A_C}}{q! \mu^q} \quad (4)$$

The originating call’ blocking probability B_{ON} is illustrated in the equation below

$$B_{ON} = \sum_{q=A_C}^A P(q) \quad (5)$$

The handoff call’ blocking probability B_{HN} is illustrated in the equation below

$$B_{HN} = P(A) = \frac{(\lambda_{ON} + \lambda_{HN})^{A_C} \lambda_{HN}^{A-A_C}}{A! \mu^A} P(0) \quad (6)$$

4.0 RESULTS AND DISCUSSION

A simulation program was developed using Java programming language. A random generator generates different data set representing phone numbers of users. The data set generated serves as the rate at which traffics enter a real life network. From the data generated, various probabilities (new call probabilities/handoff probabilities) were determined (see tables 2, 3 & 4). These probabilities were used to determine the various states of the system. The simulation parameters are shown in table 1.

Table 1: Simulation parameters

Parameter	Value
Cell capacity	1000 calls/s
Call service rate	500 calls/s
Maximum rate of generating handoff calls	600 – 800
Maximum rate of generating new calls	200 – 400
Number of sources	100 – 500
Maximum buffer size	5 - 45Mb
Transmission cycle	10 – 30

Table 2: Blocking probability against the rate of arrival at 20% new call traffic and 80% handoff call traffic.

Arrival rate	New call blocking probability	Handoff call blocking probability
0.01	0.09	0.02
0.02	0.11	0.03
0.03	0.1	0
0.04	0.05	0.02
0.05	0.08	0.01
0.06	0.02	0
0.07	0.06	0
0.08	0.05	0
0.09	0.05	0
0.10	0.06	0

Table 3: Blocking probability against the rate of arrival at 30% new call traffic and 70% handoff call traffic

Arrival rate	New call blocking probability	Handoff call blocking probability
0.01	0.06	0.01
0.02	0.05	0.01
0.03	0.11	0.01
0.04	0.06	0.01
0.05	0.06	0
0.06	0.01	0
0.07	0.06	0.01
0.08	0.11	0.03
0.09	0.06	0
0.10	0.05	0.03

Table 4: New call throughput versus handoff call throughput

Arrival rate	New call throughput	Handoff call throughput
0.01	0.99	1.00
0.02	0.95	0.96
0.03	0.88	0.98
0.04	0.99	1
0.05	0.93	0.99
0.06	0.97	0.98
0.07	0.98	1
0.08	0.97	0.99
0.09	0.93	0.98
0.10	0.94	1

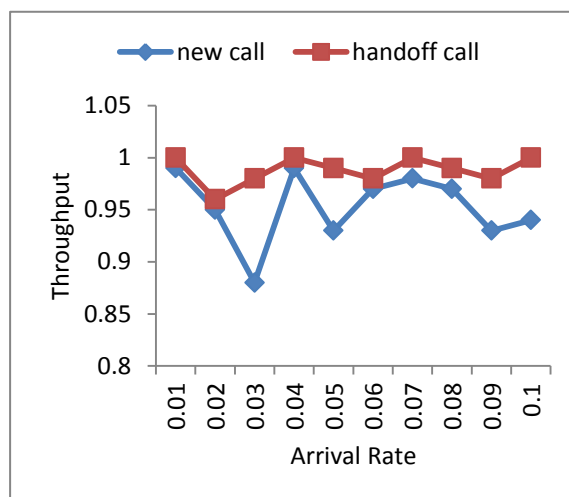


Figure 5: Blocking probability against the rate of arrival at 20% new call traffic and 80% handoff call traffic

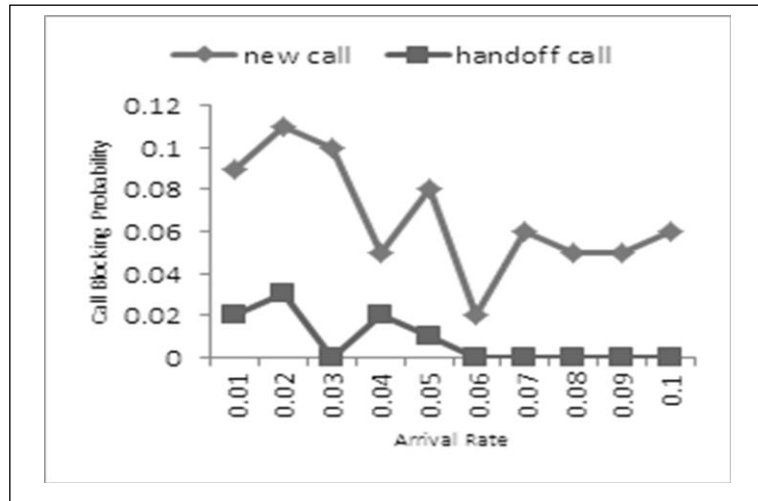


Figure 6: Blocking probability against the rate of arrival at 30% new call traffic and 70% handoff call traffic

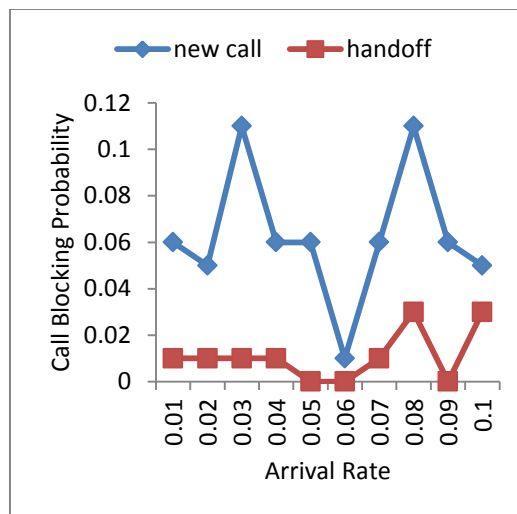


Figure 7: Throughput against the rate of arrival

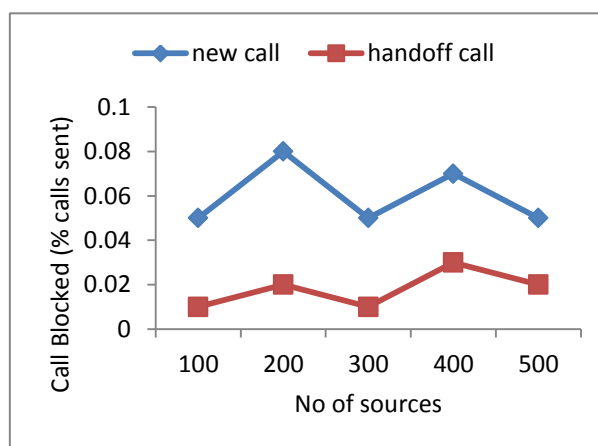


Figure 8: Blocking probability against sources number.

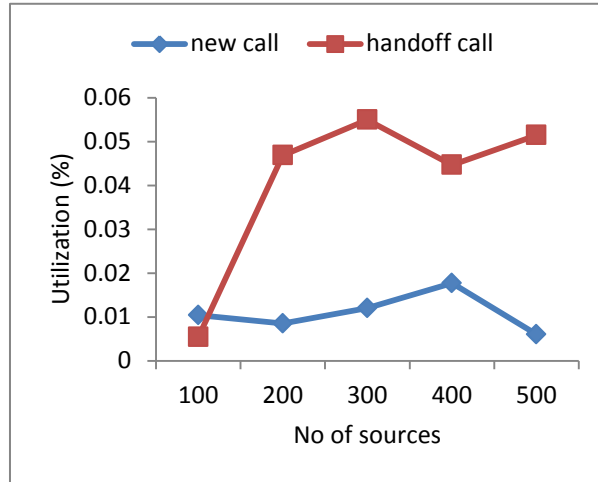


Figure 9: Resource utilization against sources number.

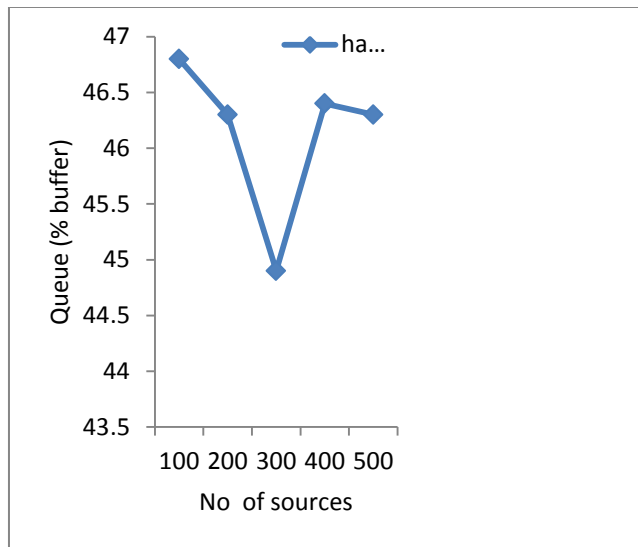


Figure 10: Percentage of buffered calls against number of sources.

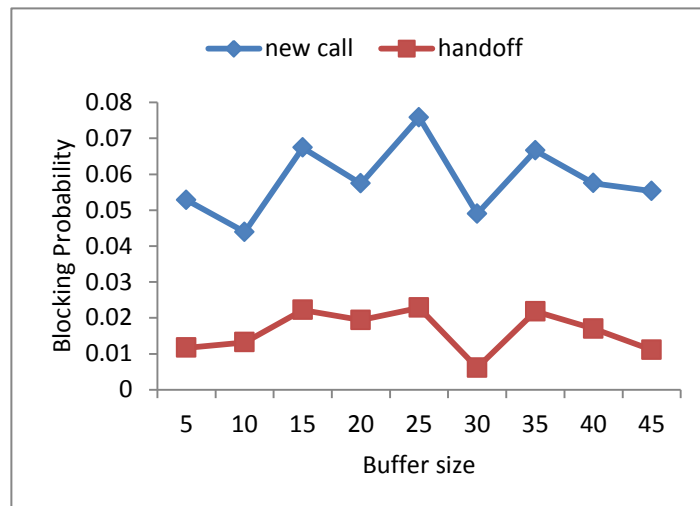


Figure 11: Call blocking probability against buffer size

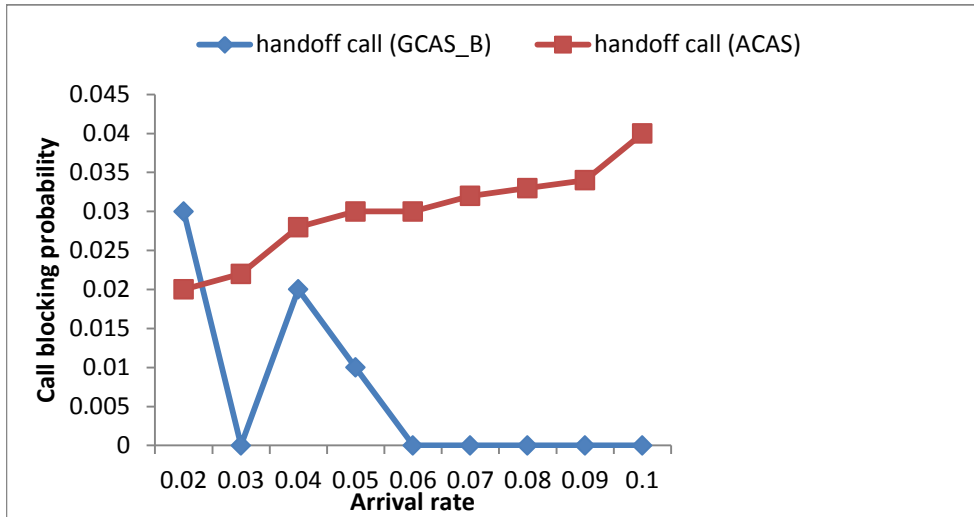


Figure 12: Comparison of handoff call blocking probability of guard channel allocation scheme with buffer (GCAS_B) with Kar and Nayak (2014) adaptive channel allocation scheme (ACAS)

Figure 5 shows the comparison of call blocking probability of new call against handoff call. The call blocking probability is against arrival rate at 20% of new call traffic and 80% handoff call. The call blocking probability for new call ranges from 0.02 to 0.11 while that of handoff call ranges from 0.00 to 0.03. It was observed that the highest call blocking probabilities for the new call and handoff call are 0.11 and 0.03 respectively. Since new call has higher call blocking probability than the handoff call, it shows that more of handoff calls are admitted. That means, at every point in time ongoing calls are given priority over new call and this is in line with many existing literatures (Alagu and Meyyappan 2012). This is an indication that handoff calls always run to completion ahead of new call due to introduction of buffer to take care of the handoff calls which otherwise would have been dropped due to unavailability of channels.

Figure 6 reveals the effect that increase in the arrival rate of new calls will have on the call blocking probability. Here the arrival rates of new call was increased to 30% while that of handoff calls was reduced to 70%. It was observed that the call blocking probability of new call was higher than that of handoff call. This is also due to the introduction of buffer to keep the blocking probability of handoff calls lower than that of new calls.

In figure 7, the comparison of throughput for both the new and handoff calls is highlighted. From the result, the throughput for the new call ranges from 0.88 to 0.99 while that of handoff call ranges from 0.96 to 1.00. It was also observed that the handoff call's highest throughput is 1.00 compared to that of new call

of 0.99. This is an indication that more handoff calls are assigned channel to run to completion and this is due to guard channels exclusively reserved for handoff calls.

Figure 8 reveals the effect of increasing the number of sources on the call blocking probability. The call blocking probability for new calls range from 0.05 to 0.08 while that of handoff calls range from 0.01 to 0.03. As the number of sources increased, the call blocking probability increased while at some times it decreased. The major reason for this instability might be because of the environment where these calls are being generated. In the overall, the handoff calls still have the lowest call blocking probability when compared with the new call. This is as a result of the introduction of buffer, and the guard channels exclusively reserved for handoff calls.

Figure 9 shows the percentage of new call and handoff call utilized. For new call, the percentage utilization ranges from 0.006 to 0.017 while that of handoff call ranges from 0.005 to 0.055. The graph clearly reveals that the handoff calls better utilized the available resources more than the new call. This showed the highest percentage utilization of 0.055 for handoff calls while that of new call is 0.017.

Figure 10 explains the effect of increasing the number of sources on the queue length. It was observed that queue reduced as the number of sources increased. Specifically, the length of the queue on the buffer increased when the number of sources increased from 300 to 400. This means more handoff calls are being generated between those sources.

Figure 11 compares call blocking probability with different buffer sizes. This is to show the most efficient buffer size that gives the lowest call blocking probability. The buffer size of 15Mb produced the highest call blocking probability while that of 30Mb produced the lowest call blocking probability. Therefore, the buffer size of 30Mb is the most efficient for the system.

Figure 12 shows the comparison of this research work, guard channel allocation scheme with buffer (GCAS_B) with the adaptive channel allocation scheme (ACAS) proposed by Kar R.R. and Nayak S.S. (2014). It was observed that the call blocking probability of handoff calls for adaptive channel allocation scheme (ACAS) was higher than that of guard channel allocation scheme with buffer (GCAS_B). The highest call blocking probability of handoff calls for guard channel allocation scheme with buffer (GCAS_B) is 0.03 while that of adaptive channel allocation scheme (ACAS) is 0.04. This shows that fewer handoff calls were dropped with guard channel allocation scheme with buffer (GCAS_B) compared to that of adaptive channel allocation scheme (ACAS) and this implies that the guard

channel allocation scheme with buffer (GCAS_B) is more efficient than adaptive channel allocation scheme (ACAS).

This research work significantly reduces the dropping of handoff calls due to the introduction of buffer which was combined with call admission control.

Parameters such as new call blocking probability, handoff call blocking probability and buffer size which were used for measuring the performance of adaptive guard channel allocation scheme with buffer were generated.

CONCLUSION

The research work addresses issue of congestion in mobile network by proposing a guard channel allocation scheme that adapts to network characteristics. Two major calls (new call and handoff call) were considered in which priority with buffer is given to handoff off calls to avoid waste of network resources. Performance measuring parameters are; blocking probability, buffer size, utilization and throughput. Further research may consider the use of double buffer.

REFERENCES

- [1] Abhinav K. and Hermant P. (2013). "A comparative study of different types of handoff strategies in cellular systems." *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, Issue 11, pp. 42-78.
- [2] Alagu, S., and Meyyappan, T. (2012a). "Efficient utilization of channels using dynamic guard channel allocation with channel borrowing strategy in handoffs." *International Conference*, New Delhi pp. 235-244.
- [3] Alagu, S. and Meyyappan, T. (2012b). "A novel adaptive channel allocation scheme to handle handoffs." *International Journal of Distributed and Parallel Systems* Vol. 3, Issue 3, pp 145-153.
- [4] Alagu, S. and Meyyappan, T. (2012c). "A novel handoff decision algorithm in call admission control strategy to utilize the scarce spectrum in wireless mobile network." *International Journal of Wireless and Mobile Networks* Vol. 4, Issue 6, p99.
- [5] Alagu, S. and Meyyappan, T. (2011d). "Analysis of handoff schemes in wireless mobile network." *International Journal of Computer Engineering Science* Vol. 1, Issue 2, pp. 1-12.
- [6] Alarape M., Akinwale A.T. and Folorunso O. (2011). "A combined scheme for controlling GSM network calls congestion." *International Journal of Computer Applications*, Vol. 14, No. 3, pp. 47-53.
- [7] Amsaveni M. and Malathy S. (2013). "Reservation of channels for handoff users in visitor location based on prediction." *Journal of Theoretical and applied Information Technology*, Vol. 57, No. 2, pp.313-318.
- [8] Biswajit B., Smita R., Parag K.G.T. and Arnab S. (2012). "Priority based hard handoff management scheme for minimizing congestion control in single traffic wireless mobile networks." *International Journal of Advancements in Technology*, Vol. 2, No 1, pp. 90-99.
- [9] Daojing, H., Caixia, C., Sammy, C., Chun, C., Jiajun, B. and Mingjian, Y. (2011). "A simple and robust vertical handoff algorithm for heterogeneous wireless mobile networks." Vol. 59, Issue 2, pp. 361-373.
- [10] Galadima A., Dajab D.D. and Bajoga G.B. (2014). "The analysis of inter cell handover dynamics in a GSM network." *International Journal of Innovative Research in Science*, Vol. 3, Issue 6, pp. 13444-13451.
- [11] Georgios I.T., Dimitrios G.S. and Erini E.T. (2010). "Call admission control in mobile and wireless networks." *National Technical University of Athens, Greece*, pp. 1-26.
- [12] Hakim, M., Alexandre, C. and Jin-Kao, H. (2011). "Genetic tabu search for robust fixed channel assignment under dynamic traffic data." Vol. 50, Issue 3, pp. 483-506.
- [13] Idil C. and Muhammed S. (2012). "Mobility based guard channel scheme for cellular networks." *Proc. Of Int. Conf. on Advances in Information and Communication*

- Technologies, pp. 51-53.
- [14] Imeh U., Oghenekaro A. and Olumide O. (2014). "Handover manageability and performance modeling in mobile communication networks." *Computing, Information Systems, Development Informatics and Allied Research Journal*, Vol. 5, No. 1, pp. 27-42.
- [15] Kar R.R. and Nayak S.S. (2014). "An efficient adaptive channel allocation scheme for cellular networks." *IOSR Journal of Computer Engineering (IOSR-JCE)*, Vol. 16, Issue 2, pp. 75-79.
- [16] Khaja K. and Aziza E. (2011), "channel assignment and minimum dropping probability scheme for handover calls in mobile wireless cellular networks." *International Journal of Recent Trends in Electrical and Electronics Engg.*, Vol. 1, Issue 2, pp. 1-9.
- [17] Kolate V. S., Patil G.I. and Bhide A.S. (2012). "Call admission control schemes and handoff prioritization in 3G wireless mobile networks." *International Journal of Engineering and Innovative Technology (IJEIT)*, Vol. 1, Issue 3
- [18] Kuboye B.M. (2010). "Optimization models for minimizing congestion in Global System for Mobile Communications (GSM) in Nigeria." *Journal Media and Communication Studies*, Vol. 2, pp. 122-126.
- [19] Lalit G., Vaibhav J. and Gourav S. (2013). "Handover management using adaptive and prioritization scheme in cellular mobile systems." *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, Issue 7, pp. 2689-2692.
- [20] Malathy S., Sudhasadasivam G., Murugan K. and Lokesh S. (2010). "Adaptive slot allocation and bandwidth sharing for prioritized handoff calls in mobile networks." *International Journal of Computer Science and Information Security*, Vol. 8, No. 1, pp. 52-57.
- [21] Ojesanmi O.A, Oyebisi T.O., Oyebode E.O. and Makinde O.E. (2011). "Performance analysis of congestion control scheme for mobile communication network." *International Journal of Computer Science and Telecommunication*, Vol. 2, Issue 8, pp. 30-37.
- [22] Raheem M.A., and Okereke O.U. (2014). "A neural network approach to GSM traffic congestion prediction." *American Journal of Engineering Research (AJER)*, Vol. 3, Issue 11, pp. 131-138.
- [23] Rejoy G., Osinanoh G.A., and Muhammad A.I. (2011). "Hybrid spectrum allocation scheme in wireless cellular networks", India-UK Advanced Technology Centre of Excellence in Next Generation Networks, Centre for Communication Systems Research, University of Surrey, Guildford, United Kingdom.
- [24] Sharif A. (2013). "Queuing-based dynamic multi-guard channel scheme for voice/data integrated cellular wireless networks." *Masters Thesis, Computer Engineering Department, Eastern Mediterranean University, Turkey*, pp. 1-66.
- [25] Shekhar V. and Geetam S.T. (2011). "Call admission control and handoff techniques for 3-G and beyond mobile networks." *Asia Pacific Journal of Multimedia Services Convergences with Art. Humanities and Sociology*, Vol. 1, No. 1, pp. 31-42.
- [26] Solomon T.G., Dominic B.O.K. and Edward N.N. (2014). "Fuzzy logic based traffic balancing in a GSM network", *Journal of Research in Engineering*, pp. 63-74.
- [27] Swati S. and Sedamkar R.R. (2012). "Improved channel assignment scheme in cellular mobile communication." *International Journal of Emerging Trends and Technology in computer science (IJETICS)*, Vol. 3, Issue 3, pp. 186-193.
- [28] Venkatachalam K. and Balasubramanie P. (2010). "Resource management for multimedia handoff calls in wireless mobile networks." *European Journal of Scientific Research*, Vol. 45, No. 2, pp. 190-199.
- [29] Vishnu K.S. and Sarita S.B. (2012). "Performance analysis on mobile agent based congestion control using AODV routing protocol technique with hop by hop algorithm for mobile ad-hoc network." *International Journal of Ad hoc, Sensor and Ubiquitous Computing (IJASUC)*, Vol. 3, No. 2, pp. 49-64.
- [30] Xin W. Arunita J. and Ataul B. (2011). "Optimal channel allocation with dynamic power control in cellular networks." *International Journal of Computer Networks and Communications (IJCNC)*, Vol. 3, No. 2, pp. 83-93.

THE EVALUATION OF TERTIARY INSTITUTION SERVICE QUALITY USING HiEdQUAL AND FUZZY TOPSIS

O. O. Oladipupo, T. O. Amoo and O. J. Daramola

*Department of Computer Science, College of Science and Technology,
Covenant University, Ota, Ogun State, Nigeria*

*funke.oladipupo@covenantuniversity.edu.ng, taiwo.amoo@stu.cu.edu.ng,
olawande.daramola@covenantuniversity.edu.ng*

ABSTRACT

One of the most important decisions that affect the future of young students is a decision as regards a Tertiary Institution of choice. In making such a decision, a number of factors are required which include service quality. Service quality consists of different attributes and many of them are intangible and difficult to measure, which means that using the conventional measurement approach is insufficient. This study presents an effective approach for evaluating and comparing service qualities of four Higher Institutions. Fuzzy set theory is adopted as a research template to resolve the ambiguity of service quality concepts and capture intra-uncertainties, which are associated with human judgments in decision making. In this study Extended HiEdQUAL educational service quality model was adopted to evaluate the respondents' judgments of service quality, Multi Attribute Decision Making method: TOPSIS is applied for the comparison among the tertiary Institutions. The importance weight of performance criteria are determined with Fuzzy Analytical Hierarchy Process (FAHP). All the algorithms were implemented using Java programming language. This study was able to present the importance of each service quality factor, quantitatively reveal each institution's weak and strong points, and rank the institutions according to the multiple criteria service quality measure.

Keywords: Institutions, Fuzzy MCDM, Quality of Services, Ranking, Decision Making

1.0 INTRODUCTION

In driving the economic growth and sustainability of a nation, tertiary institutions are pivotal. Higher education was pronounced to be more resourceful to the nations' growth than primary or secondary education [1]. As a follow up to this in Nigeria, tuition fee for Federal and State Higher institutions are subsidized to encourage more students. Also, federal and state universities receive external funding from the government in order to provide quality education to their primary customers. This external funding is supposed to be the impetus for driving service quality in the institutions by seeing to the delivery and increasing satisfaction of demands of the

primary customers. This should in turn lead to an increase in clients' preferences, creation of more values and heightened excellence [2]. However institutions in Nigeria are yet to attain expected global excellence and relevance. This has been confirmed with unfulfilled expectations from their primary customers [2]. This is further depicted in their unsatisfactory performance in the global ranking of universities.

In [3] service quality was defined as the capacity of an organization to equal or surpass the customers' expectations. According to [4], the quality dispensation of services has always been the differentiator

that represents a university's prestige among her contemporaries and the high-level make-up of its nations' human capital. In [5], it was observed that students prefer to choose a university that has evidence of quality of service delivery. It is usually quite tasking when the students have to do an evaluation in order to select the best institutions from a set of alternative institution. In such cases, it is important to sort, describe or do a ranking of the alternatives in order to make a good decision or recommendation. In this situation, the Multi Criteria Decision Making (MCDM) methods become useful. MCDM refers to making decisions in the presence of multiple, and conflicting criteria [6]. MCDM can be broadly classified into 2 categories which are the Multi Attribute Decision Making (MADM) and Multi-Objective Decision Making (MODM) [7,8]. MADM belongs to a class of methods that solve decision making problems that are discrete in nature i.e. have finite number of alternatives to be evaluated while the MODM approach, on the other hand, encompasses methods that deal with decision making problems that are non-deterministic in nature, whereby the decision space is continuous and alternatives are infinite [8,9,10].

The evaluation of quality of service (QoS) in higher institutions is a Multi-Criteria Decision Making (MCDM) problem due to its multi-criteria nature. Hence, a multi-dimensional classification of the criteria is needed for the evaluation of higher institutions quality of service. MCDM methods have capability to accommodate the variations in the notion of each decision maker's (DM) representation of order of preference or importance of the criteria that drives the perception of service quality. The MCDM accesses each alternative versus the criteria through qualitative and quantitative measurements thereby giving an overall utility value for each alternative and ranking them from the best to the least according to the decision maker's opinions.

In the real world, opinions differ and consist of uncertainties when measurement are

made under human consideration. This is not any different in the evaluation of service quality. Therefore, in order to handle the subjectivity and uncertainty in the opinions of decision makers, fuzzy set theory becomes essential for the MCDM process for modeling the decision makers' opinions. The incorporation of Fuzzy Set theory allows the use of linguistic terms like *Fair, Strong and Very Strong* and the like, and membership values in measuring the satisfaction level anticipated for each criterion with respect to the alternative concerned.

So far, in the measurement of service quality, different models have been developed. SERVQUAL [3] is a service quality model with widespread applicability in various service industries [11]. It sought out to obtain the gap between customers' expectation minus the perception of customers' service outcome quality. The measurement scale includes reliability, assurance, responsiveness, tangibility, and empathy. Due to the complexities and intangible nature of services in the educational sector, it has been argued that SERVQUAL might not be sufficient to handle variations in the educational sector [11,12]. Therefore, in order to cater for all dimensions and variables related to the educational sector, the HiEdQUAL as a service quality model was proposed in [12,5], which sought to embrace key items and dimensions that are necessary to evaluate service quality in higher institutions. Therefore, in this study, the HiEdQUAL service quality model has been adopted with additional dimensions of INTERNATIONALIZATION to support the world ranking target criteria.

Since the measurement of service quality could help an organization to be positioned strategically in order to maintain a competitive edge over the competitors, the emphasis of evaluating service quality in a higher institution cannot be over emphasized. Hence the motivation for computational evaluation of quality of service in four

Nigeria tertiary institutions (two private and 2 public universities). This study used fuzzy MCDM approach in order to compare the universities, underlying quality of service delivery dimension from the perspective of their administrators and students. Also, this study assessed the qualitative judgments of the administrators and students of the institutions with respect to their preferences in reaching a consensus.

The remaining part of this paper are as follows. In Section 2, an extensive literature review was carried out. In Section 3, the methodology was explicitly described while in Section 4, the result was shown and discussed in Section 5. The paper is concluded in Section 6 with a summary and overview of future works.

2.0 Related works

According to [13], all MCDM methods evaluating alternatives using numerical analysis have these three steps: (i) determine the criteria and alternatives; (ii) determine the weights for each criterion to show order of importance and the scores of each criterion with respect to the alternative; and (iii) process the numerical values to aid in ranking of the alternatives. MCDM techniques have the capacity for alternatives to be measured explicitly through objective and subjective judgements of decision maker. Each method is only unique in how it combines its data and thereby give different ranking results [13,14]]. MCDM has the capability to accommodate both quantitative and qualitative measurement of criteria. This has made it suitable for evaluating service quality in different sectors, including educational sector, especially when the performance evaluation problem is qualitative in nature.

Researchers have contributed immensely in using different MCDM methods for the evaluation of alternatives [15,16,17,18,19]. Fuzzy MCDM models have been reported as widely used approaches in decision making processes [16]. In [18] a hybrid MCDM method for performance evaluation of private universities in Taiwan was presented. The

performance evaluation indices used in the work as a benchmark, was based on an official performance evaluation structure developed by the Taiwan Assessment and Evaluation Association (TWAEA). AHP was used to weigh the performance evaluation indices and the VIKOR method in ranking the private universities. However, the work was structured to be geographically context specific to universities in Taiwan. A web based support system using the MCDM method was developed with ELECTRE III in the personalized ranking of British Universities [19]. In [20], VIKOR was employed as MCDM method in the ranking of universities in Turkey based on academic performance only. In [21], a study was presented by utilizing the AHP and TOPSIS in the evaluation of performance of schools. Parents were utilized as the determinants of the performance of each criterion. This ranking was not based on higher institutions but on high schools. TOPSIS with fuzzy type-2 was used in [22], for managing the choice of a university. The service quality was based on SERVQUAL model, which is too general to handle variations in the educational sector for ranking Higher Institutions [12,13].

Therefore, in this study a context-specific scientific evaluation index/model is considered in the evaluation of Quality of Service (QoS) in Tertiary Institutions. HiEdQUAL as a higher institution domain specific service quality dimension of criteria is extended and used for ranking some tertiary institutions in Nigeria. Fuzzy analytical hierarchy process (FAHP) is used to obtain criteria weight and TOPSIS for ranking the cases. The Fuzzy concept was introduced to resolve the ambiguity of the concepts and capture intra-uncertainty, which are associated with the decision maker's judgments. The ranking order preference is by similarity to ideal solution. FAHP Eigen vector algorithm using the iterative Power method and The TOPSIS algorithm were implemented with Java programming Language. This study was able to present the general importance of each service quality

factors, reveal each institution's weak and strong points quantitatively and rank the institutions accordingly based on multiple criteria.

3.0 Methodology

In this work, the evaluation process is distinctly covered in 4 steps as shown in Fig. 1, which are i) definition and establishment

of criteria to evaluate the higher institutions based on their quality of service; ii) the determination of each criterion weight; iii) determination of criteria performance for each alternative; and iv) evaluation and ranking of alternatives. The alternatives considered for evaluation are 2 private universities and 2 public universities from the south-west region of Nigeria.

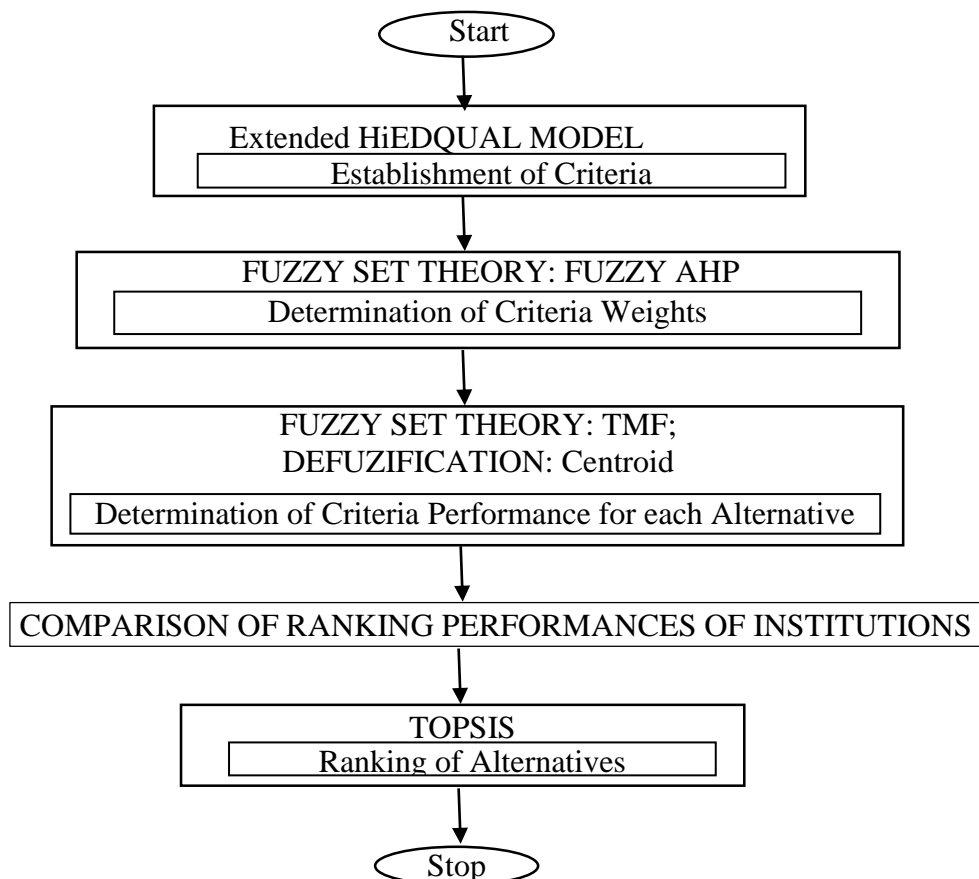


Figure1: The Tertiary Institution service quality assessment workflow

3.1 Criteria Definition

In this study, HiEdQUAL service quality model which is specific with respect to measurement of service quality in higher institutions was adopted with few additional criteria. The HiEdQUAL model is based on 5 dimensional concepts which are Teaching and Course content (TC), Administrative Services (AS), Academic Facilities (AF), Campus Infrastructure (CI) and Support Service (SS). Each dimensional concept has its attributes embedded such that they are weighted individually. To fully come to terms with

Expectations of the students in relation to QoS and evaluate the global relevance of Nigerian universities, another factor/criterion- INTERNALISATION (IN)- was considered. This additional criterion underlines the importance of fast tracking excellence in higher institutions and attaining global relevance. The additional attributes which make up IN that were considered include:

- i) *University provides international exchange programs*

- ii) University has a number of international lecturers/faculty;
- iii) University has standard collaborations for recruitment of international staff and students; and
- iv) University has international students.

These attributes can provide the focal point for rapid development of institution’s by-products, acceleration of its goals and global

However, such attributes were not sufficiently embedded in the HiEDQUAL model. The extended HiEDQUAL service quality model produced 6 concepts and 33 criteria. Each of the criteria is rated based on the fuzzy linguistic values {*very dissatisfied, dissatisfied, fair, satisfied, and very satisfied*}. The extended HiEDQUAL QoS criteria model and its notations are shown in Table 1.

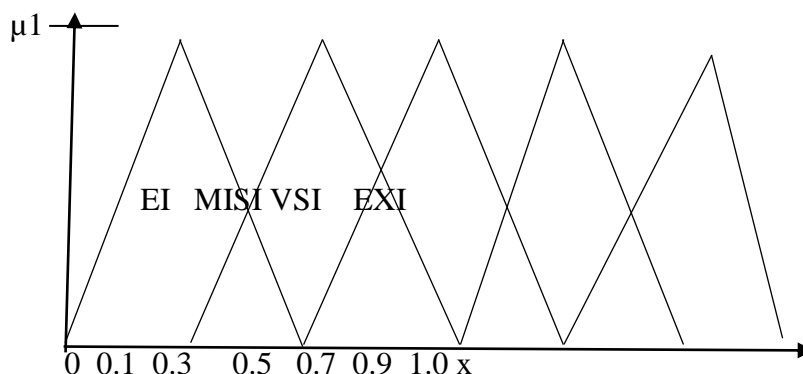


Figure 2: Fuzzy sets of the linguistic values for Perceived importance

Key: EI- Equally important, MI- Moderately more important, SI- Strongly more important, VSI- Very strongly more important, EXI-Extremely more important.

relevance among its contemporaries.

3.2.1 Fuzzification Process

This models the uncertainties and imprecision of the decision makers involved in the evaluation process. The students are the decision makers, being the primary stakeholders of tertiary institutions. In order to elicit criteria importance and institutions performance with respect to the criteria, 2 linguistic variables are defined and fuzzified based on literature [6]: the “Perceived Importance” and “Perceived Performance”. These linguistic variables represent the judgements of the decision makers for eliciting the perceived importance of one criterion over another and perceived performance of each alternative in relation to extended HiEdQUAL QoS model. The linguistic values for the linguistic variable “Perceived Importance” for eliciting each criterion importance from decision makers are identified as:

moderately important, strongly important, very strongly important, and extremely important

Each fuzzy set is represented by the fuzzy graph in Figure. 2.

Also, the linguistic values for the linguistic variable “Perceived Performance” for eliciting the performance of each institution with respect to the QoS criteria is defined as Perceived Performance {*very dissatisfied, dissatisfied, fair, satisfied, very satisfied*}.

Each fuzzy set is represented by the fuzzy graph in Figure 3.

The membership function $\mu_A(x)$ of a Triangular Fuzzy Number (TFN) is defined in (1):

Perceived Importance {*equally important,*

Table 1: The Extended HiEDQUAL criteria model with corresponding weights

The Evaluation of Tertiary Institution Service Quality Using HiEdQUAL and Fuzzy Topsis
O.O. Oladipupo, T.O. Amoo and O.J. Daramola

S/N	CRITERIA CODE	SELECTED CRITERIA	WEIGHT
1.	TC1	Teachers are responsive and accessible	0.0273
2.	TC2	Teachers follow curriculum strictly	0.0306
3.	TC3	Teachers follow good teaching practices	0.0292
4.	TC4	Relevance between programme & syllabus	0.0182
5.	TC5	Course content develops students' knowledge	0.0262
6.	TC6	Department Informs schedules, exams, results on time	0.0270
7.	TC7	Teachers Complete syllabus on time	0.0312
8.	TC8	Department has sufficient academic staff	0.0350
9.	TC9	University has more adjunct lecturers than in-house lecturers	0.0159
10.	TC10	Departments reflect current trends in the curriculum	0.0166
11.	AS1	Administrative staff provide service without delay	0.0278
12.	AS2	Administrative staff are courteous and willing to help	0.0298
13.	AS3	Administrative staff provide error free work	0.0312
14.	AS4	Administration maintains accurate and retrieval records	0.0361
15.	AS5	Administrative staff are accessible during office hours	0.0161
16.	AS6	University has safety and security measures	0.0217
17.	AF1	Departments have adequate teaching facilities	0.0438
18.	AF2	Classrooms equipped with teaching aids	0.0518
19.	AF3	Department has sufficient class rooms	0.0206
20.	AF4	University has adequate auditoriums, conference halls etc.	0.0314
21.	AF5	Library has adequate academic resources	0.0387
22.	AF6	Computer labs have adequate equipment and internet facilities	0.0326
23.	C11	University has adequate hostel facilities	0.0341
24.	C12	University has adequate medical facilities (Health centres)	0.0311
25.	C13	University has adequate social amenities (Canteen, Shopping Mall, Bank, ATM, Post office, etc.)	0.0394
26.	C14	Campus infrastructure is well maintained.	0.0334
27.	SS1	University has sufficient sports and recreation facilities.	0.0273
28.	SS2	University/department provides placement services.	0.0306
29.	SS3	University provides counselling services	0.0292
30.	IN1	University provides international exchange programmes /collaboration	0.0182
31.	IN2	University has a number of international lecturers/faculty.	0.0262
32.	IN3	University has standard collaborations for recruitment of international staff and students.	0.0270
33.	IN4	University has international students.	0.0312

$$(\mu_x)Triangle = \left\{ \begin{array}{l} 0 \text{ if } x \leq k \text{ or } x > m \\ \frac{x-k}{l-k} \text{ if } k < x \leq l \\ \frac{m-x}{m-l} \text{ if } l < x \leq m \end{array} \right\} \quad (1)$$

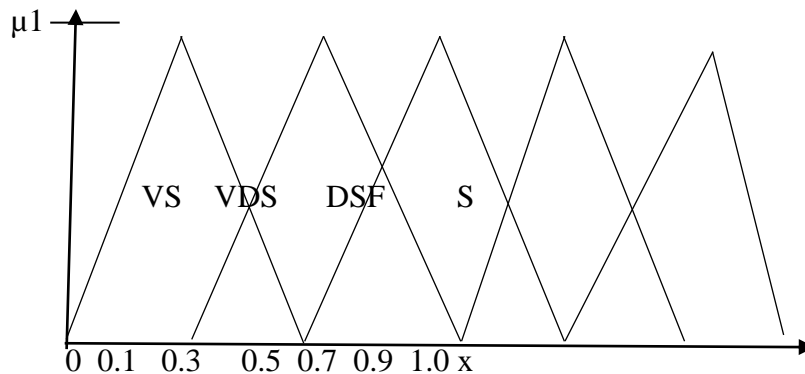
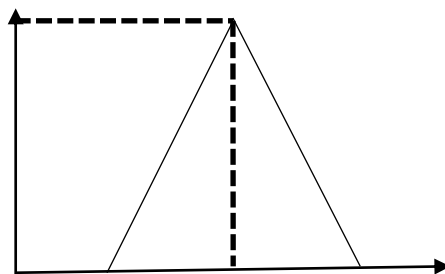


Figure 3: Fuzzy sets of the linguistic values for Perceived performance

Key: VDS = very dissatisfied, DS = dissatisfied, F = fair, S = satisfied, VS = very satisfied

where: k is lower boundary; l is median point and m is upper boundary, $m \neq l$, $u \neq m$ as depicted in Fig. 4 such that $\mu_A(x): X \rightarrow \{0,1\}$, where

$$\begin{aligned} \mu_A(x) &= 1 && \text{if } x \text{ is totally in } A; \\ \mu_A(x) &= 0 && \text{if } x \text{ is not in } A; \\ &(2) \\ 0 < \mu_A(x) < 1 && \text{if } x \text{ is partially in } A \end{aligned}$$



3.2.2 The Fuzzy Aggregation Process

Based on the inherent subjectivity in the evaluation process, the perceptions of the decision makers are aggregated using the arithmetic operations of fuzzy set defined in equation (3). 50 students in each of the four institutions were polled using a random survey in relation to the extended HiEdQUAL model to determine the

performance of the institutions respectively. These perceptions are in two parts: the perceptions based on criteria importance (Perceived Importance) and secondly, on alternative performance with respect to the HiEdQUAL criteria Model (Perceived Performance). Consequently, these perceptions modelled as fuzzy variables, are averaged using the following formula in equation (4) which is expounded in [22,23] as:

$$A_{ave} = \frac{1}{n} \left(\sum_{i=1}^n k_i, \sum_{i=1}^n l_i, \sum_{i=1}^n m_i \right) \quad (3)$$

Its average triangular fuzzy number [31] is now:

$$\begin{aligned} A_{ave} &= (k_a, l_a, m_a) \\ &= \left(\frac{1}{n} \sum_{i=1}^n k_i, \frac{1}{n} \sum_{i=1}^n l_i, \frac{1}{n} \sum_{i=1}^n m_i \right) \end{aligned} \quad (4)$$

Equation (4), gives us an average fuzzy quantitative performance value for each criterion with respect to the decision makers' assessment. The Fuzzy performance for each Higher Institution (HI) with respect to the criteria is shown in Table 2.

Table 2: Fuzzy performance for each Higher Institution (HI) with respect to the modified HiEdQUAL. A,B,C,D represent the Four Universities under consideration

Criteria	Fuzzy Criteria Performance of the Higher Institutions(HI) (Fuzzification)			
	HI A	HI B	HI C	HI D
TC1	(3.600,5.480,7.400)	(5.040,7.040,8.680)	(4.580,6.560,8.380)	(4.600,6.600,8.480)
TC2	(4.120,6.120,8.020)	(5.240,7.240,8.900)	(4.660,6.640,8.480)	(4.200,6.200,8.140)
TC3	(3.640,5.560,7.460)	(4.740,6.720,8.500)	(3.920,5.920,7.860)	(4.040,6.040,7.980)
TC4	(3.800,5.720,7.640)	(4.580,6.520,8.360)	(3.740,5.720,7.620)	(4.560,6.560,8.420)
TC5	(3.820,5.720,7.580)	(5.040,7.000,8.760)	(4.460,6.440,8.280)	(4.680,6.680,8.540)
TC6	(4.240,6.160,7.980)	(5.240,7.240,8.800)	(3.540,5.480,7.380)	(3.760,5.760,7.720)
TC7	(3.800,5.800,7.740)	(4.620,6.600,8.320)	(2.820,4.720,6.700)	(3.360,5.360,7.340)
TC8	(3.600,5.520,7.460)	(4.780,6.720,8.420)	(3.460,5.400,7.300)	(4.440,6.440,8.360)
TC9	(3.240,5.160,7.120)	(3.540,5.440,7.320)	(3.140,5.080,7.000)	(2.780,4.760,6.720)
TC10	(3.140,5.040,7.040)	(4.560,6.520,8.280)	(3.420,5.360,7.320)	(3.800,5.800,7.780)
AS1	(2.880,4.760,6.720)	(3.900,5.840,7.800)	(2.720,4.640,6.580)	(2.940,4.880,6.840)
AS2	(3.600,5.520,7.460)	(4.240,6.200,8.040)	(3.040,5.000,6.940)	(3.340,5.240,7.220)
AS3	(3.240,5.120,7.060)	(3.600,5.560,7.420)	(3.040,5.040,7.020)	(3.280,5.240,7.200)
AS4	(3.820,5.720,7.600)	(3.920,5.880,7.780)	(3.900,5.840,7.780)	(3.600,5.600,7.560)
AS5	(3.840,5.760,7.600)	(4.240,6.200,8.060)	(4.040,6.040,7.940)	(4.400,6.400,8.300)
AS6	(3.340,5.160,7.100)	(5.060,7.040,8.660)	(4.240,6.240,8.060)	(3.620,5.600,7.560)
AF1	(2.620,4.400,6.340)	(4.040,6.000,7.840)	(2.620,4.480,6.400)	(3.360,5.280,7.220)
AF2	(2.340,4.120,6.100)	(4.760,6.760,8.500)	(2.540,4.440,6.360)	(3.600,5.560,7.500)
AF3	(3.420,5.360,7.340)	(4.160,6.080,7.860)	(2.880,4.720,6.680)	(4.360,6.320,8.200)
AF4	(3.040,4.920,6.840)	(4.860,6.840,8.520)	(3.480,5.400,7.280)	(4.020,5.960,7.860)
AF5	(4.120,6.120,7.940)	(5.640,7.640,9.160)	(3.820,5.800,7.600)	(3.940,5.920,7.800)

AF6	(2.640,4.440,6.400)	(4.820,6.760,8.420)	(3.020,4.840,6.740)	(2.780,4.680,6.640)
C11	(2.140,3.880,5.860)	(4.600,6.560,8.300)	(2.760,4.640,6.620)	(1.140,2.600,4.580)
C12	(2.540,4.320,6.300)	(4.680,6.640,8.360)	(3.100,5.040,6.940)	(3.080,5.040,7.000)
C13	(2.920,4.760,6.720)	(4.580,6.520,8.240)	(2.660,4.560,6.500)	(4.320,6.320,8.240)
C14	(2.380,4.120,6.100)	(4.080,6.040,7.920)	(3.620,5.560,7.440)	(2.760,4.720,6.680)
SS1	(3.000,4.840,6.780)	(4.060,6.000,7.820)	(3.340,5.280,7.200)	(3.740,5.720,7.620)
SS2	(2.640,4.440,6.420)	(3.200,5.120,7.000)	(3.040,4.920,6.860)	(2.800,4.720,6.700)
SS3	(2.240,4.000,6.000)	(4.340,6.280,8.060)	(3.780,5.720,7.540)	(3.580,5.520,7.480)
IN1	(2.180,3.880,5.880)	(4.400,6.360,8.200)	(3.040,4.960,6.920)	(2.580,4.440,6.420)
IN2	(1.285,2.836,4.836)	(3.560,5.520,7.380)	(2.220,4.000,5.980)	(2.100,3.920,5.900)
IN3	(1.775,3.571,5.571)	(3.880,5.800,7.680)	(2.380,4.160,6.140)	(2.320,4.240,6.220)
IN4	(1.632,3.285,5.285)	(3.600,5.520,7.340)	(2.260,4.080,6.040)	(2.280,4.160,6.100)

3.2.3 Defuzzification

According to [24], excessive fuzzification does not infer better modelling of reality, this could end up being counter-productive. There is a need to transform the fuzzy nature of both the average performance values of each criterion and the criteria importance values into best non-fuzzy value since final judgements are made with crisp values.

Therefore, in order to transform it into its best non-fuzzy performance value, the centroid defuzzification method was employed in this study because it is monotonous, consistent, and its deterministic response curve is

Characterized by a smooth and continuous behavior [25]. The defuzzified value of a fuzzy number can be attained as in (5). The

Best Non-Fuzzy Performance values using centroid defuzzification with respect to selected criteria is shown in Table 3.

$$BNP = \frac{(m_i - k_i) + (l_i - k_i)}{3} + k_i \quad (5)$$

Where,

k_i represents the first aggregated triangular fuzzy number for the overall performance of criterion i for HI_j

l_i represents the middle aggregated triangular fuzzy number for the overall performance of criterion i for HI_j

m_i represents the third aggregated triangular fuzzy number for the overall performance of criterion i for HI_j

Table 3:BNP for each Higher Institution (HI) with respect to the criteria

Criteria	Best Non-Fuzzy Performance of the Higher Institutions(HI) (Centroid defuzzification)			
	HI A	HI B	HI C	HI D
AS1	4.7867	6.0600	4.6467	4.8867
AS2	5.5267	4.4933	4.9933	5.2667

AS3	5.1400	3.9600	5.0333	5.2400
AS4	5.7133	4.3867	5.8400	5.5867
AS5	5.7333	4.8000	6.0067	6.3667
AS6	5.2000	4.2000	6.1800	5.5933
AF1	4.4533	4.8733	4.5000	5.2867
AF2	4.1867	4.5000	4.4467	5.5533
AF3	5.3733	4.0800	4.7600	6.2933
AF4	4.9333	3.9800	5.3867	5.9467
AS1	4.7867	6.0600	4.6467	4.8867
AF5	6.0600	2.9533	5.7400	5.8867
AF6	4.4933	3.5933	4.8667	4.7000
C11	3.9600	3.3600	4.6733	2.7733
C12	4.3867	4.7867	5.0267	5.0400
C13	4.8000	5.5267	4.5733	6.2933
C14	4.2000	5.1400	5.5400	4.7200
SS1	4.8733	5.7133	5.2733	5.6933
SS2	4.5000	5.7333	4.9400	4.7400
SS3	4.0800	5.2000	5.6800	5.5267
IN1	3.9800	4.4533	4.9733	4.4800
IN2	2.9533	4.1867	4.0667	3.9733
IN3	3.5933	5.3733	4.2267	4.2600
IN4	3.3600	4.9333	4.1267	4.1800

3.3 FuzzyAnalytical Hierarchy Process (FAHP)

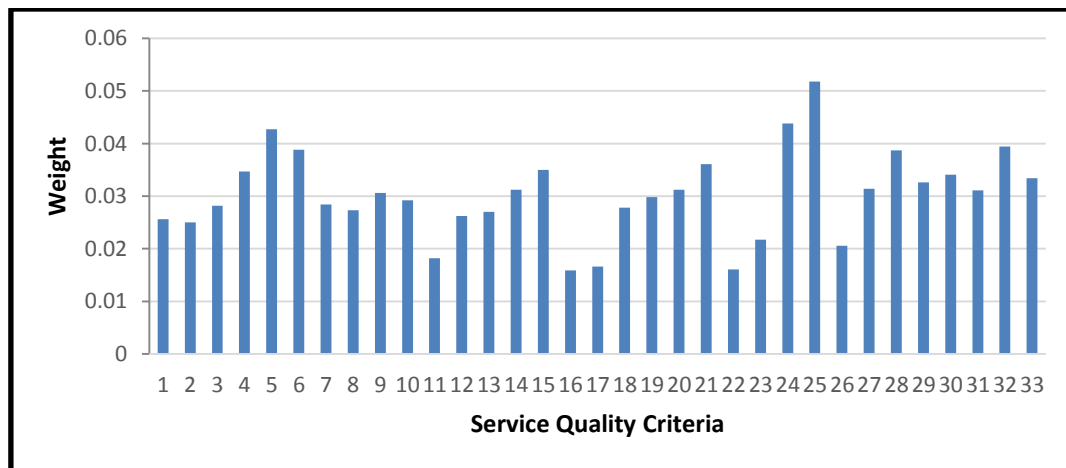


Fig. 5: Graphical representation of the Fuzzy AHP derived weights of extended HiEDQUAL QoS criteria

AHP was proposed by Saaty [26], and is one of the most popularly utilized weighting MCDM method [15]. The fusion of Fuzzy set concept addresses the weakness of the AHP by capturing

AHP was proposed by Saaty [26], and is one of the most popularly utilized weighting MCDM method [15]. The fusion of Fuzzy

set concept addresses the weakness of the AHP by capturing the uncertainties of the decision makers before making recommendations on criteria importance. Fuzzy AHP was adopted as a weighing model for determining each criterion weight as shown in Table 1 and graphically represent in Figure 5. The Experts

considered in this study are people in the University Quality Assurance unit, Academic Planning and Head of Departments from the universities under consideration due to their deeper knowledge about their universities' quality of service. The experts are considered for the rating of the criteria importance based on the fuzzy linguistic values {*Equally important, moderately more important, strongly more important, very strongly more important, extremely more important*}.

For the determination of weights/priorities, using Fuzzy AHP approach, Eigen vector algorithm, using the iterative Power method has been adopted [27], and implemented in

$$\begin{pmatrix} 1 & (\tilde{a}_{12k}, \tilde{a}_{12l}, \tilde{a}_{12m}) \dots (\tilde{a}_{1nk}, \tilde{a}_{1nl}, \tilde{a}_{1nm}) \\ (\tilde{a}_{21k}, \tilde{a}_{21l}, \tilde{a}_{21m}) & 1 & \dots (\tilde{a}_{2nk}, \tilde{a}_{2nl}, \tilde{a}_{2nm}) \\ \vdots & \vdots & \ddots & \vdots \\ (\tilde{a}_{n1k}, \tilde{a}_{n1l}, \tilde{a}_{n1m}) & (\tilde{a}_{n2k}, \tilde{a}_{n2l}, \tilde{a}_{n2m}) \dots & \dots & 1 \end{pmatrix}$$

Java Programming language due to its ability to accommodate slight inconsistencies in the comparisons matrix of the high number of criteria being considered. The Fuzzy AHP algorithm is stated as:

Step 1: Construct a fuzzy pairwise comparison matrix for each decision maker (DM) with the aid of the linguistic scale in Figure 2. Each element, (\tilde{a}_{ij}) in the pairwise comparison matrix, \tilde{A} is a fuzzy number corresponding to the linguistic value selected by the decision maker (DM) in Figure 2. Consequently, the relative importance of one criterion over another can be subjectively expressed by a DM in constructing the pairwise comparison matrix using the following steps:

- If two criteria have equal importance in pairwise comparison, enter corresponding fuzzy number in Figure 2 for both criteria;
- If one of them is moderately more important than the other, enter the corresponding fuzzy number and for the other enter reciprocal fuzzy number in Figure 2;
- If one of them is strongly more important than the other, enter the

corresponding fuzzy number and for the other enter reciprocal fuzzy number in Figure 2;

- If one of them is very strongly more important than the other, enter the corresponding fuzzy number and for the other enter reciprocal fuzzy number in Figure 2;
- If one of them is extremely important than the other, enter the corresponding fuzzy number and for the other enter reciprocal fuzzy number in Figure 2;

Then, we have:

$$\tilde{A} = \begin{pmatrix} 1 & \tilde{a}_{12} & \dots & \tilde{a}_{1n} \\ \tilde{a}_{21} & 1 & \dots & \tilde{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 1 \end{pmatrix} \text{ rewritten}$$

as

(6)

where $(a_{12k}, a_{12l}, a_{12m})$ is represented as triangular fuzzy number for criterion i and j and

$(a_{21k}, a_{21l}, a_{21m})$ represents the reciprocal fuzzy number for the comparison between criterion j and i

Step 2: Add the fuzzy numbers selected by each DM for each pairwise comparison of ith and jth criterion using the equation (3) if there is more than one DM, and collapse into an aggregated one using equation (4).

Step 3: Defuzzify averaged fuzzy numbers from step (2) for each $(a_{1nk}, a_{1nl}, a_{1nm})$ in \tilde{A} using the Centroid

Defuzzification formula in equation (5) due to the condition required of linear additive models like AHP as certainty is prerequisite before final results.

Step 4: Square the averaged Pairwise comparison matrix: $A_{n+1} = A_n * A_n$.

Step 5: Perform the row sums that are calculated and normalized using equation (7) and (8):

$$ri = \sum_i aij \quad (7)$$

summation of the row values and

$$pi = \frac{ri}{\sum_i ri} \quad (8)$$

where π_i is the normalization of the sums. This produces the first eigenvector.

Step 6: Repeat steps 4 and 5 using the new matrix A_{n+2} derived from squaring the A_n

Step 7: Stop when there is a difference between the current and last eigenvector solution in two consecutive priorities calculations derived from Step 5 and 6.

3.4 Ranking of Alternatives Using TOPSIS MCDM Methods

For the ranking of the tertiary institutions of learning under consideration, TOPSIS MCDM method was applied. TOPSIS proposed by Hwang and Yoon [6] was developed based on the idea that the best alternative should have the shortest distance to the positive ideal situation and the farthest to the negative ideal situation.

Step 1: Calculation of the normalized performance matrix

$$r_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^m a_{ij}^2}} \quad i = 1, 2, \dots, m \quad j = 1, 2, \dots, n \quad (9)$$

where a_{ij} is the performance matrix, i as the alternative number and j the criterion number

Step 2: Calculation of weighted and normalized performance matrix

$$V_{ij} = w_j \times r_{ij} \quad i = 1, 2, \dots, m \quad j = 1, 2, \dots, n \quad (10)$$

where, w_j is weight of criterion j .

Step 3: Determination of positive ideal solution and negative ideal solution.

The weighted and normalized values in matrix V_{ij} gives rise to the positive ideal solution (A^+) and Negative ideal solution (A^-)

$$A^+ = (V_1^+, V_2^+ \dots V_n^+) \quad (11)$$

$$A^- = (V_1^-, V_2^- \dots V_n^-) \quad (12)$$

the V_{ij} matrix, V_j^- is the worst alternative value in criterion j from the V_{ij} matrix.

The Normalized Performance matrix and Weighted Normalized Performance matrix for each Higher Institution (HI), with respect to the criteria is shown on Table 4

Step 4: Calculation of distance values between alternatives

The distance of alternatives from the positive ideal solution S_i^+ and the distance of alternatives from the negative ideal solution S_i^- are calculated with Eqs. (13) and (14) respectively

$$S_i^+ = \sqrt{\sum_{j=1}^n (V_{ij} - V_j^+)^2} \quad i = 1, 2, \dots, m \quad (13)$$

$$S_i^- = \sqrt{\sum_{j=1}^n (V_{ij} - V_j^-)^2} \quad i = 1, 2, \dots, m \quad (14)$$

Step 5: Calculation of the closeness to the positive-ideal solution

$$C_i^+ = \frac{S_i^-}{S_i^+ + S_i^-} \quad i = 1, 2, \dots, m \quad (15)$$

C_i^+ value is in the $0 \leq C_i^+ \leq 1$ interval. As C_i^+ gets closer to 1, alternative i gets closer to A^+ , whereas if C_i^+ gets closer to 0, alternative i gets closer to A^- .

Step 6: Arrangement of alternative choices: Alternatives are arranged according to the decreasing order of C_i^+ .

4.0 Results

The result from this study reveal some hidden and non-trivial knowledge about the four universities considered as related to the QoS criteria measures. Table 1 shows the 33 criteria with their corresponding weight which reveals the criteria importance as related to the institution under consideration. This was determined by FAHP Eigen vector algorithm, using the iterative Power method. The graphical representation is shown in Figure 5. Each institution performance against each criterion is tabulated in Table 3. Table 4 shows each institution performance

positive and negative closeness to the ideal performance as against each criterion using TOPSIS MCDM method. For the final ranking of the institutions' performance with TOPSIS, the result of the closeness to the positive-ideal solution associated with each alternative is shown in **Table 5** and graphically represented on Figure 6. Figures 7-10 show the graphical weight of each criterion for the four universities considered.

method

Alternative Institution	TOPSIS	Ranking
Higher Institution A	0.1595	4
Higher Institution B	0.9702	1
Higher Institution C	0.3416	3
Higher Institution D	0.4307	2

Table 5: Ranking values for the TOPSIS

Selected Criteria	Fuzzy Criteria Performance of the Higher Institutions(HI) (Fuzzification)			
	HI A	HI B	HI C	HI D
TC1	(3.600,5.480,7.400)	(5.040,7.040,8.680)	(4.580,6.560,8.380)	(4.600,6.600,8.480)
TC2	(4.120,6.120,8.020)	(5.240,7.240,8.900)	(4.660,6.640,8.480)	(4.200,6.200,8.140)
TC3	(3.640,5.560,7.460)	(4.740,6.720,8.500)	(3.920,5.920,7.860)	(4.040,6.040,7.980)
TC4	(3.800,5.720,7.640)	(4.580,6.520,8.360)	(3.740,5.720,7.620)	(4.560,6.560,8.420)
TC5	(3.820,5.720,7.580)	(5.040,7.000,8.760)	(4.460,6.440,8.280)	(4.680,6.680,8.540)
TC6	(4.240,6.160,7.980)	(5.240,7.240,8.800)	(3.540,5.480,7.380)	(3.760,5.760,7.720)
TC7	(3.800,5.800,7.740)	(4.620,6.600,8.320)	(2.820,4.720,6.700)	(3.360,5.360,7.340)
TC8	(3.600,5.520,7.460)	(4.780,6.720,8.420)	(3.460,5.400,7.300)	(4.440,6.440,8.360)
TC9	(3.240,5.160,7.120)	(3.540,5.440,7.320)	(3.140,5.080,7.000)	(2.780,4.760,6.720)
TC10	(3.140,5.040,7.040)	(4.560,6.520,8.280)	(3.420,5.360,7.320)	(3.800,5.800,7.780)
AS1	(2.880,4.760,6.720)	(3.900,5.840,7.800)	(2.720,4.640,6.580)	(2.940,4.880,6.840)
AS2	(3.600,5.520,7.460)	(4.240,6.200,8.040)	(3.040,5.000,6.940)	(3.340,5.240,7.220)
AS3	(3.240,5.120,7.060)	(3.600,5.560,7.420)	(3.040,5.040,7.020)	(3.280,5.240,7.200)
AS4	(3.820,5.720,7.600)	(3.920,5.880,7.780)	(3.900,5.840,7.780)	(3.600,5.600,7.560)
AS5	(3.840,5.760,7.600)	(4.240,6.200,8.060)	(4.040,6.040,7.940)	(4.400,6.400,8.300)
AS6	(3.340,5.160,7.100)	(5.060,7.040,8.660)	(4.240,6.240,8.060)	(3.620,5.600,7.560)

AF1	(2.620,4.400,6.340)	(4.040,6.000,7.840)	(2.620,4.480,6.400)	(3.360,5.280,7.220)
AF2	(2.340,4.120,6.100)	(4.760,6.760,8.500)	(2.540,4.440,6.360)	(3.600,5.560,7.500)
AF3	(3.420,5.360,7.340)	(4.160,6.080,7.860)	(2.880,4.720,6.680)	(4.360,6.320,8.200)
AF4	(3.040,4.920,6.840)	(4.860,6.840,8.520)	(3.480,5.400,7.280)	(4.020,5.960,7.860)
AF5	(4.120,6.120,7.940)	(5.640,7.640,9.160)	(3.820,5.800,7.600)	(3.940,5.920,7.800)
AF6	(2.640,4.440,6.400)	(4.820,6.760,8.420)	(3.020,4.840,6.740)	(2.780,4.680,6.640)
C11	(2.140,3.880,5.860)	(4.600,6.560,8.300)	(2.760,4.640,6.620)	(1.140,2.600,4.580)
C12	(2.540,4.320,6.300)	(4.680,6.640,8.360)	(3.100,5.040,6.940)	(3.080,5.040,7.000)
C13	(2.920,4.760,6.720)	(4.580,6.520,8.240)	(2.660,4.560,6.500)	(4.320,6.320,8.240)
C14	(2.380,4.120,6.100)	(4.080,6.040,7.920)	(3.620,5.560,7.440)	(2.760,4.720,6.680)
SS1	(3.000,4.840,6.780)	(4.060,6.000,7.820)	(3.340,5.280,7.200)	(3.740,5.720,7.620)
SS2	(2.640,4.440,6.420)	(3.200,5.120,7.000)	(3.040,4.920,6.860)	(2.800,4.720,6.700)
SS3	(2.240,4.000,6.000)	(4.340,6.280,8.060)	(3.780,5.720,7.540)	(3.580,5.520,7.480)
IN1	(2.180,3.880,5.880)	(4.400,6.360,8.200)	(3.040,4.960,6.920)	(2.580,4.440,6.420)
IN2	(1.285,2.836,4.836)	(3.560,5.520,7.380)	(2.220,4.000,5.980)	(2.100,3.920,5.900)
IN3	(1.775,3.571,5.571)	(3.880,5.800,7.680)	(2.380,4.160,6.140)	(2.320,4.240,6.220)
IN4	(1.632,3.285,5.285)	(3.600,5.520,7.340)	(2.260,4.080,6.040)	(2.280,4.160,6.100)

5.0 DISCUSSION

In this paper, we have presented an empirical study on ranking of four universities (2 private and 2 public) in the south-west region of Nigeria based on fuzzy TOPSIS MCDM method. The adopted HiEdQUAL model of five

Table 4: The Normalized Performance matrix and Weighted Normalized Performance matrix for each Course content (T), Administrative

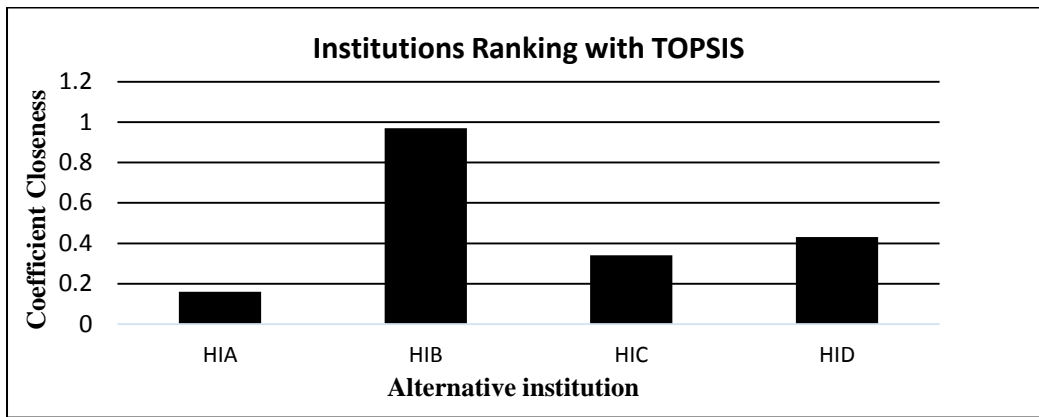


Figure 6: Graphical representation of the

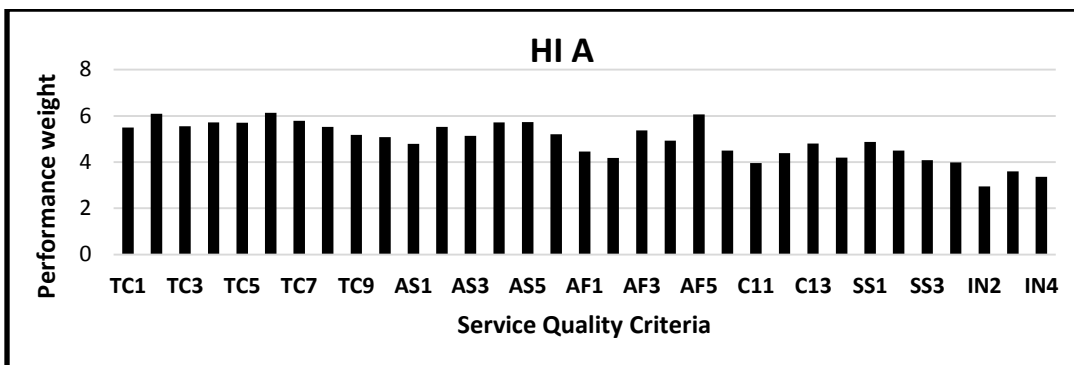


Figure 7: Higher Institution 'A' performance against each criterion



Figure 8: Higher Institution 'B' performance against each criterion

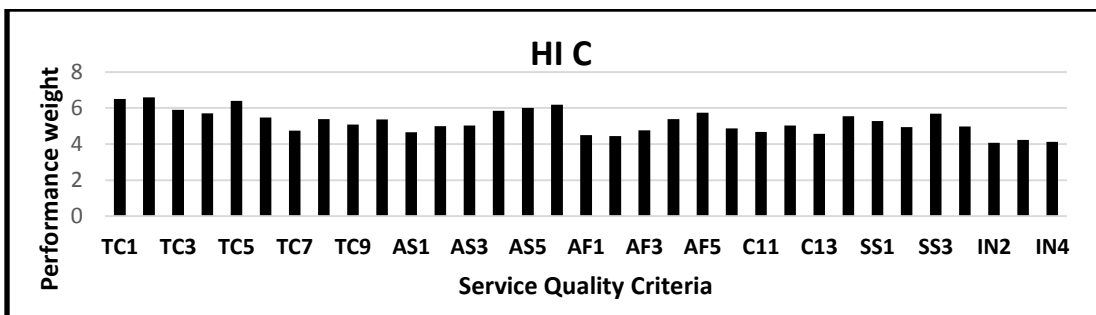


Figure 9 Higher Institution 'C' performance against each criterion

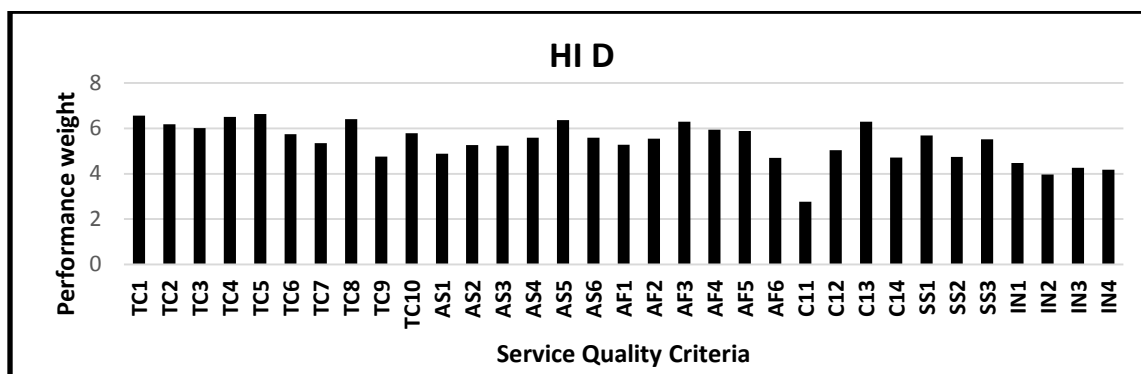


Figure 10: Higher Institution ‘D’ performance against each criterion

This is considered based on one-one interaction with the decision makers. This amount to 33 criteria for the evaluation process as opposed to other models in [18,19,20,21,22] which is relatively specific to cater for demands of the universities in their respective geographical regions. The additional dimension reveals the international exposure and collaboration of the institutions. For example a criterion like *recruitment of international staff* can fast track realignment of the university goals in repositioning their policies for global ranking if the weakness is sufficiently exposed to the university.

As depicted in Figure5, it is observed that Criterion 25 (*University has adequate Amenities - Canteen, Shopping center, Bank, ATM, Post office etc.*) has the highest weight, followed by Criterion 24 (*University has adequate medical facilities (Health centers)*). The lowest weight was Criterion 16 (*University has safety and security measures*). This indicates that in ranking universities in Nigeria, the stakeholder are more concerned about the amenities and facilities as compared to safety and security. Meaning that they are less concern about the safe and security the universities can guarantee.

From Table 4, the performance weight of each alternative institution against each criterion is determined. This actually dictates to the institutions their strong fits of quality of service and the weak points. Figures. 7 and 8 give a graphical picture of alternative HIA and HIB performance against each

criterion respectively. From Figure.7, HIA has the strongest weight in TC6 with 6.13 performance weight and the weakest weight in IN2 with 2.95. This shows that HIA has to improve on the number of international lecturers/faculty and maintain her strength in prompt release of Departmental schedules, exams and results which is their main strength for now. For HIB from Figure 7, the strongest point is AS1 with 6.06 weight. This shows that the institution HIB’s strength is in services provided by the administrative staff promptly without delay. Therefore, there could be need to motivate their administrative staff to do more since this is her major strength. The weakest fit is AF5 with 2.95. This suggests that there is need to improve in her Library academic resources to enhance her quality of service.

Also, according to Table 5, the alternative HIB is evaluated and ranked the best alternative with the 0.97 coefficient of closeness to the ideal. This was followed by HID with 0.43 coefficient of closeness, HIC with 0.34 and HID with 0.16 coefficient of closeness. From Figure 5, it is clearly shown that higher institution B is the best ranked institution based on the 33 criteria. This means HIB is the closest institution to the Ideal Institution among others, followed by institution D. The farthest institution away from the Ideal institution is institution A. Nevertheless, HIB has its own weak point as shown in Figure 8 and Alternative HIA being the least ranked institution, also has its own strong point as shown in Figure 7. This implied that being the best is relative

per time. There is always a room for improvement to satisfy the primary stake holders.

The results obtained could be more generalizable if more participants are involved as against the 50 students employed to determine the performance of their respective universities. Meanwhile, this study gives a pointer to administrators in their respective universities to addressing the concerns of their primary stakeholders (students) in terms of service delivery. The result from this study can mitigate the effect of slow development and serve as an eye opener for institution in repositioning for global ranking of world class universities.

6.0 CONCLUSION

REFERENCES

1. W. Ho, , P.K. Dey and H.E. Higson, Multiple criteria decision-making techniques in higher education, *International journal of educational management*, 20(5), 2006,319-337.
2. R. Asiyai, Challenges of quality in higher education in Nigeria in 21st century. *International Journal of Educational Planning & Administration*,3(2), 2013, 159-172.
3. A Parasuraman, VA. Zeithaml, LL. Berry, A conceptual model of service quality and its implications for future research, *the Journal of Marketing*, 1985,41-50.
4. R. Nazarian , Saber-Mahani, M. and Beheshtifar, M., Role of Service Quality in Universities,*Innova Ciencia*. 4(6), 2012, 3-9
5. B. Donaldson, C. McNicholas Understanding the postgraduate education market for UK-based students: a review and empirical study. *International Journal of Non-profit and Voluntary Sector Marketing*, 9(4), 2004,346-60.
6. C. Hwang, K. Yoon, *Multiple Attribute Decision Making: Methods and Application* (Springer-Verlag, New York,1981).
7. EK. Zavadskas, Z. Turskis, S. Kildienė, State of art surveys of overviews on MCDM/MADM methods, *Technological and economic development of economy*, 20(1), 2014,165-179.
8. E. Mulliner, N. Malys, V. Maliene, Comparative analysis of MCDM methods for the assessment of sustainable housing affordability,*Omega*,59(2), 2016,146-156.
9. GH. Tzeng, JJ. Huang, *Fuzzy multiple objective decision making* (CRC Press, 2013).
10. J. Lu, Zhang, GH. Tzeng, JJ. Huang, *Fuzzy multiple objective decision making* (CRC Press; 2013).
11. G. Ruan , F. Wu *Multi-Objective Group Decision Making: Methods, Software and Applications with Fuzzy Set Techniques (With CD-ROM)*, (World Scientific, 2007).
12. F. Abdullah The development of HEDPERF: a new measuring instrument of service quality for the higher education sector, *International Journal of Consumer Studies*, 30 (6), 2006; 569-581.

13. S. Annamdevula, RS. Bellamkonda Development of HiEdQUAL for Measuring Service Quality in Indian Higher Education Sector. *International Journal of Innovation, Management and Technology*,3(4), 2012;412.
14. E. Triantaphyllou, *Multi-criteria decision making methods: a comparative study* (Springer Science & Business Media, 2000).
15. J. Antucheviciene, A Zakarevicius, EK. Zavadskas Measuring congruence of ranking results applying particular MCDM methods, *Informatica*. 22(3), 2011,319-38.
16. H. Akdag, T. Kalaycı, S. Karagöz, H. Zülfikar, D. Giz. The evaluation of hospital service quality by fuzzy MCDM, *Applied Soft Computing*,23, 2014, 239-48.
17. A Mardani, A. Jusoh, EK. Zavadskas, Fuzzy multiple criteria decision-making techniques and applications—Two decades review from 1994 to 2014. *Expert Systems with Applications*,42(8),2015,4126-48.
18. HY. Wu, JK. Chen, IS. Chen, HH. Zhuo, Ranking universities based on performance evaluation by a hybrid MCDM model. *Measurement*, 45(5), 2012,856-80.
19. C. Giannoulis, A. Ishizaka, A Web-based decision support system with ELECTRE III for a personalised ranking of British universities. *Decision Support Systems*. 48(3), 2010,488-9
20. S. Nisel, R. Nisel, Using VIKOR methodology for ranking universities by academic performance, Proc. International Conf. on Operations Research and Statistics (ORS), Global Science and Technology Forum, 2013.
21. JM. Vinotha. Analytic hierarchy process and TOPSIS method to evaluate the performance of schools. *International Journal in IT & Engineering*,3(3), 2015,29-37.
22. M Erdoğan, İ. Kaya, A type-2 fuzzy MCDM method for ranking private universities in İstanbul, Proc. of the World Congress on Engineering 2014, 2-4.
23. Dubois D, Prade H. Operations on fuzzy numbers. *International Journal of systems science*, 9(6), 1978, 613-26.
24. JJ. Buckley, Ranking alternatives using fuzzy numbers, *Fuzzy sets and systems*, 15(1), 1985,21-31.
25. RA. Ribeiro, Fuzzy multiple attribute decision making: a review and new preference elicitation techniques. *Fuzzy sets and systems*. 78(2), 1996,155-181.
26. DT. Pham, M. Castellani, Action aggregation and defuzzification in Mamdani-type fuzzy systems. Proc. of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 2002,216(7),747-759.
27. T.L. Saaty, *The Analytic Hierarchy Process: Planning, Priority Setting* (McGrawHill, New York,1980).

CyberProtector: IDENTIFYING COMPROMISED URLs IN ELECTRONIC MAILs WITH BAYESIAN CLASSIFICATION

N.A Azeez¹ and O. Ademolu²

^{1,2}*Department of Computer Sciences, University of Lagos, Nigeria.*

¹*nazeez@unilag.edu.ng; ²oluwatosin.ademolu@yahoo.com*

ABSTRACT

Embedding malicious URLs in e-mails is one of the most common web threats facing the internet community today. Malicious URLs have been widely used to mount various cyber-attacks like spear phishing, pharming, phishing and malware. By falsely claiming to be a trustworthy entity, users are lured into clicking on these compromised links to divulge vital information such as usernames, passwords, or credit card details and unknowingly succumb to identity theft. Hence, the detection of malicious URLs in e-mails is very essential so as to help internet users implement safe practices and as well prevent them from becoming victims of fraud. This paper explores how malicious links in e-mails can be detected from the lexical and host-based features of their URLs to protect users from identity theft attacks. This research uses Naïve Bayesian classifier as a probabilistic model to detect if a URL is malicious or legitimate. The Naïve Bayesian classifier is used to count up the occurrence of each feature in an e-mail and calculate the cumulative score. If the cumulative score is greater than the given threshold, the URL is considered malicious otherwise the URL is legitimate.

Keywords: Malicious URLs; Pharming; Phishing, Attacks; Naïve Bayesian classifier, threshold.

I. INTRODUCTION

Although the World Wide Web is a very fast and easy tool for sharing information over the internet, it also has an immense risk of cyber-attacks. Scammers have used the Web as a means of delivering malicious attacks such as phishing, pharming, e-mail spoofing and malware infection [1]. For instance, phishing which is a form of identity theft involves sending an e-mail that appears to be from a trustworthy source to lure people into clicking a compromised Uniform Resource Locator (URL) contained in the e-mail which links to a fake website in an attempt to illegally get a user's vital personal information such as usernames, passwords, credit card details, social security or bank account numbers [15].

According to the Anti-Phishing Working Group (APWG), Phishing Activity Trends

Report for the fourth quarter of 2015, the total number of unique phishing websites detected was 158,574 while the total number of unique phishing e-mail reports received by APWG from consumers was 380,280. It was estimated that 70% of internet users have received phishing e-mails, out of which approximately 15% have been lured into providing their personal information which is subsequently used for fraudulent purposes [3].

Spoofed e-mails which contain malicious URLs are sent to users by posing to be a legitimate entity asking for urgent response to conditions such as account suspension, failed transaction or upgrade to a newly installed security feature in order to get sensitive information. These URLs once clicked, lead to the actual phishing sites which are replicas of

genuine websites and lure the users into entering sensitive information [4].

Many users are not skilled enough to defend themselves against identity theft and fraud attacks because the URLs of phishing websites are forged to look very similar, and sometimes even identical, to the legitimate websites. So it is difficult for even a more careful user to detect fraudulent websites.

There is therefore, a need for an efficient method to defend users against such attacks [14].

Figure 1 is a sample phishing e-mail. In an effort to circumvent this great challenge, the authors propose a project titled “Identifying Compromised URLs in Electronic Mail with Bayesian Classification”.



Figure 1: Sample Phishing E-mail

II. RELATED WORKS

Over the past few years, several methods have been applied to detect and defend against phishing attacks. Here, we briefly review some existing anti-phishing techniques.

Zhang, et al. (2007) presented a content-based approach for detecting phishing websites using an anti-phishing tool called Carnegie Mellon Anti-phishing and Network Analysis tool (CANTINA). Unlike other heuristic-based phishing detection approaches that examine the visible characteristics of a web page (e.g the URL and its domain name) to classify a website as either phishing or legitimate, the CANTINA was designed to examine the content of a web page to determine whether or not a website is legitimate. CANTINA was also combined with some heuristics like checking the age of the domain name, checking if a page contains inconsistent well-known logos or images, checking if the URL and links of a page are suspicious, checking if

a page's domain name is an IP address, checking the number of dots in a page's URL, checking if a page contains forms with submit button. This approach is however cumbersome and extremely difficult to implement [24].

Ram, et al. (2008) talked about the effectiveness of using different machine learning algorithms for the classification of phishing emails. They compared the performance of six different machine learning methods in detecting and classifying phishing emails which include Support Vector Machines (SVMs), Biased Support Vector Machines (BSVMs), A Library for Support Vector Machines (LIBSVM), Neural Networks (NN), K-Means, Self Organizing Maps (SOMs). They found that LIBSVM achieved consistently the best results.

Medvet, et al. (2008) presented an anti-phishing system that detects phishing attempts by comparing the visual similarity between a suspected phishing site and the legitimate site

that is spoofed. They reported that it was typical of victims to judge and be convinced of a webpage’s authenticity by its look-and-feel. Therefore, three features that are visually perceived by users were considered to determine page similarity: the texts & style of text in the webpage, the images embedded in the webpage, and the overall visual appearance of the webpage as rendered by the browser. A webpage is reported as a phishing site if its similarity to the legitimate webpage is higher than the predefined similarity threshold [21].

Xiang & Hong (2009) presented a hybrid phishing detection approach based on information extraction (IE) and information retrieval (IR) techniques. Their method used an identity-based detection component to detect phishing sites by discovering the inconsistencies between a website’s real identity and its claimed identity and a keywords-retrieval detection component which employs IR algorithms and exploits the power of search engines to detect phishing. These two

components manipulate the Document Object Model (DOM) after the webpage has been rendered in a web browser to bypass deliberate obstacles [22].

Han, et al. (2012) proposed an Automated Individual White-List (AIWL) approach to protect user’s web digital identities. AIWL detects web digital identities theft attacks such as phishing and pharming by keeping an automated individual white-list of all web sites familiar to the user together with the Login User Interface information of these websites [8]. A Naïve Bayesian classifier was used by the AIWL to automatically build an individual white-list for a user. If a user tries to submit his or her account information to a website that does not match the white-list, AIWL will alert the user of the possible attack. AIWL also keeps track of the features of login pages (e.g., IP addresses, document object model (DOM) paths of input widgets) in the individual white-list and checks the legitimacy of these features to defend users against attacks [6].

Table 1: Common Phishing Features [2]

S.No.	Phishing Features	No. of appearances	Appearance %
1	Using the IP address	14	46.66 ←
2	Abnormal request URL	30	100 ←
3	Abnormal URL of anchor	7	23.33 ←
4	Abnormal DNS record	2	06.66
5	Abnormal URL	5	16.66
6	Using SSL certificate	17	56.66 ←
7	Certification authority	4	13.33
8	Abnormal cookie	2	06.66
9	Distinguished Names Certificate (DN)	4	13.33
10	Redirect pages	3	10.00
11	Straddling attack	2	06.66
12	Pharming attack	4	13.33
13	Using on MouseOver to hide the link	6	20.00 ←
14	Server Form Handler (SFH)	2	06.66
15	Spelling errors	24	80.00 ←
16	Copying website	5	16.66
17	Using forms with “Submit” button	6	20.00 ←
18	Using Pop-Ups windows	8	26.66 ←
19	Disabling right click	2	06.66
20	Long URL address	22	73.33 ←
21	Replacing similar characters for URL	16	53.33 ←
22	Adding prefix or suffix	9	30.00

23	Using the @ symbol to confuse	6	20.00 ←
24	Using hexadecimal character codes	8	26.66 ←
25	Much emphasis on security and response	5	16.66
26	Buying time to access accounts	3	10.00

III. PHISHING FEATURES

There are several phishing features, however, based on the review of various related researches, the feature set has been reduced and the common phishing features which will help to improve the accuracy and the precision of detecting malicious URLs were gathered as shown in the Table 1.

According to the analysis by (Aburrou, Hossain, Keshav, & Fadi, 2010), the above underlined features showed high impact in various studies as mentioned in the related works and hence for better performance, the feature set comprises features whose impact is greater than 20%. This includes the host based features, lexical features, and suspicious keywords in the e-mail [2] [19].

URL FEATURES USED

Phishing URLs can be examined based on two types of features: lexical features and host-based features of the URL [5]. The lexical features analyse the format of the URL while the host based features identify the location, owner and how malicious sites are hosted and managed.

Lexical Features

Lexical features are the textual properties of the URL. It analyses the format of the URL not the content of the page it references. These properties include the length of the entire URL, presence of IP address in URL, the number of dots in the URL, presence of phishing keywords in URL, presence of suspicious characters such as @ symbol, hexadecimal characters and use of delimiters or special binary characters like “/”, “?”, “:”, “=”, “-”, “\$”, “^” either in the host name or path [5]. It should be noted that F1 to F8 are the features considered in this work.

- a. *Length of URL (F1)*: Most phishing URLs use very large domain names to lure end-users so that the URL may appear legitimate. e.g. *http://www.tsv1899benningen-ringen.de/chronik/update/alert/ibclogon.php*. Thus, if the length of a URL is longer than 55 characters, the URL is flagged suspicious.
- b. *Use of IP address in URL (F2)*: Some phishing websites contain an IP address in their URL instead of the domain name in order to hide the actual domain name which is malicious. When the URL in an email has its host name as an IP address. For example, in *http://65.222.204.76/co/*, we flag the URL suspicious.
- c. *Using the hexadecimal character codes (F3)*: A malicious URL can also be represented using hexadecimal base values with a ‘%’ symbol to hide the actual letters and numbers in the URL. Thus, a URL that has hexadecimal character codes will be flagged suspicious [11].
- d. *Use of @ symbol in URL (F4)*: The ‘@’ character is used by phishers to make host names difficult to understand. A @ symbol in a URL will enable the string to the left of the ‘@’ symbol which is the actual legitimate URL to be discarded while the string to the right which leads to the phishing site is treated as the actual website. For example, in the URL *http://www.worldbank.com@phishingsite.com*, “*www.worldbank.com*” will be discarded while “*phishingsite.com*” will be treated as the actual domain name. When a URL contains the ‘@’ symbol is detected, we flag it suspicious.

- e. *Number of dots in URL (F5)*: Usually, legitimate URLs will not contain more than five dots but phishing URLs typically have many dots because phishers make use of sub domains to make the URLs look legitimate. Having sub domains means having an extremely large number of dots in the URL. For example, “<https://login.personal.wamu.com/verification.asp?d=1>” has two sub domains. A URL with more than five (5) dots is flagged as phishing [10].
- f. *Number of Sensitive Words in URL (F6)*: Some sensitive words frequently appear in phishing URLs such as secure, account, update, login, sign-in, banking, confirm, verify suspend, username, etc. URLs that contain one or more of these keywords are deemed suspicious. We use this feature to flag a URL as phishing.

Host-Based Features

Host-based features describe the location of malicious sites, that is, where they are being hosted, who these sites are managed by and how they are managed. Some of these features are age of domain, page rank, number of domains [7].

- g. *Age of domain (F7)*: The age of the domain identifies when a website is hosted such that a website that has less age or is relatively new is flagged suspicious. Many phishing sites have registered domain names that exist only for a short period of time to evade detection. They may be recently registered and some domains may not even be available at the time of checking. The WHOIS lookups on the WHOIS server is used to retrieve the domain registration date, and if the domain registration entry is not found on the WHOIS server, the URL is considered suspicious [9].
- h. *Presence of Form Tag (F8)*: One of the methods phishers use to collect information from users is the use of form tag in URL. For example, `<FORM`

`action=http://www.paypalsite.com/profile.php method=post`, the PayPal URL contains a form tag which has the action attribute actually sending the information to `http://www.paypalsite.com/profile.php` and not to `http://www.paypal.com`. Thus, a URL that has the form tag is flagged suspicious.

- i. *Number of Domains (F9)*: A phishing URL may contain two or more domain names which are used to forward address from one domain to the other. For example, “`http://www.google.com/url?sa=t&ct=res&cd=3&url=http%3A%2F%2Fwww.antiphishing.org%2F&ei=-0qHRbWHK4z6oQLTmBM&usg=uIZX_3aJvESkMveh4uItI5DDUzM=&sig2=AVrQFpFvihFnLjpnGHVsxQ`” has two domain names where “google.com” forwards the click to “antiphishing.org” domain name. The number of domain names in the URL extracted from an e-mail is counted and if more than one, we flag the URL suspicious [12].
- j. *HTTP Status of URL (F10)*: Some URLs are not valid at all. A URL is checked and if the URL passes the HTTP ok code 200, it means the URL is valid. If the HTTP ok code 200 returns as false then the URL is not valid and is flagged phishing.

IV. NAÏVE BAYESIAN (NB) CLASSIFIER

Naïve Bayesian (NB) classification algorithm is based on applying Bayes’ theorem with strong (naive) independence assumptions between the features. Naïve Bayesian classifier is one of the most successful learning algorithms for text categorisation, that is, grouping documents as belonging to one category or the other (such as spam or legitimate) with word frequencies as the features and it is widely used in Spam Filtering [20].

Given a dependent class variable C with a small number of outcomes or classes which is conditional on several feature variables, each URL in an email is represented by a feature

vector $\vec{F} = (F_1, F_2, F_3, \dots, F_n)$ where each of the property, $F_1, F_2, F_3, \dots, F_n$ is independent. A Naive Bayes classifier can be represented as follows:

$$P(C = c|F_1, \dots, F_n) = \frac{P(C = c).P(F_1, \dots, F_n|C = c)}{\sum_{k \in (\text{spam}, \text{legitimate})} P(C = k).P(F_1, \dots, F_n|C = k)} \dots \dots (1)$$

The ‘‘naive’’ conditional independence assumes that each feature F_i is conditionally independent of every other feature F_j ($j \neq i$) given a class C . Hence, $P(C = c|F_1, \dots, F_n)$ can be computed as:

$$P(C = c|F_1, \dots, F_n) = \frac{P(C = c). \prod_{i=1}^n P(F_i|C = c)}{\sum_{k \in (\text{spam}, \text{legitimate})} P(C = k). \prod_{i=1}^n P(F_i|C = k)} \dots \dots \dots (2)$$

where $P(F_i|C)$ and $P(C)$ can be easily calculated from the training samples.

V. INTEGRATING BAYESIAN PROBABILITY WITH THE PROPOSED SYSTEM

Using Bayes’ theorem, the conditional probability for the independent feature vector $\vec{F} = (F_1, F_2, F_3, \dots, F_n)$ with k possible outcomes or classes C_k is:

$$P(C_k|F) = \frac{P(C_k)P(F|C_k)}{P(F)} \dots \dots \dots (3)$$

That is,
 posterior = $\frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \dots \dots \dots (4)$

Since URLs can be classified as valid or malicious, new URLs are classified as they arrive, that is, the class label they belong to is determined based on the currently existing URLs. Prior probabilities are based on previous experiences, in this case the percentage of Valid and Malicious URLs are used to predict outcomes before they actually happen.

$$\text{Prior Probability of Valid URLs} = \frac{\text{Number of Valid URLs}}{\text{Total number of Tested URLs}} \dots \dots \dots (5)$$

$$\text{Prior Probability of Malicious URLs} = \frac{\text{Number of Malicious URLs}}{\text{Total number of Tested URLs}} \dots \dots \dots (6)$$

Since there is a total of (X) Tested URLs, (Y) of which are Valid URLs and (Z) Malicious URLs, the prior probabilities for class membership are:

$$\text{Prior Probability of Valid URLs} = \frac{Y}{X} \dots \dots \dots (7)$$

$$\text{Prior Probability of Valid URLs} = \frac{Z}{X} \dots \dots \dots (8)$$

Having formulated the prior probability, a new URL can now be classified. A new URL (N) can be classified using the likelihood probability:

$$\text{Likelihood of N given a Valid URL} = \frac{\text{Number of Valid URLs}}{\text{Number of Tested Valid URLs}} \dots \dots \dots (9)$$

$$\text{Likelihood of N given a Malicious URL} = \frac{\text{Number of Malicious URLs}}{\text{Number of Tested Malicious URLs}} \dots \dots \dots (10)$$

The final classification is done by combining both sources of information, i.e., the prior and the likelihood probabilities, to form a posterior probability using Bayes' rule.

$$\begin{aligned} &\text{Posterior probability of new URL (N)being Valid} \\ &= \text{Prior Probability for Valid URLs} \\ &\times \text{Likelihood of N given a Valid URL} \dots \dots \dots (11) \end{aligned}$$

$$\begin{aligned} &\text{Posterior probability of new URL (N)being Malicious} \\ &= \text{Prior Probability for Malicious URLs} \\ &\times \text{Likelihood of N given a Malicious URL} \dots \dots \dots (12) \end{aligned}$$

ALGORITHM I: NAÏVE BAYESIAN CLASSIFIER

Input:URL

Output:Class Value

1. Extract URLs from e-mail
 - 1.1. $host=Host(URL)$
 - 1.2. $path=Path(URL)$
2. Extract URL features
 - 2.1. $features [F_1, F_2, F_3, F_4, F_5, F_6]=Extract(host, path)$ //Each feature F_i takes value 0, or 1
3. Calculate probability of malicious URL
 - 3.1. $P(C_p|features) = P(C_p) * \prod_{i=1 to 6} P(x_i|C_p)$ // C_p is class phishing
4. Calculate probability of legitimate URL
 - 4.1. $P(C_l|features) = P(C_l) * \prod_{i=1 to 6} P(x_i|C_l)$ // C_l is class legitimate
5. if $(P(C_l|features) / P(C_p|features) > \alpha)$ Class=1 //legitimate
6. else-if $(P(C_p|features) / P(C_l|features) > \alpha)$ Class=-1 //phishing
7. else-if $((1/\alpha) < P(C_l|features) / P(C_p|features) < \alpha)$ Class=0 //suspicious

VI. PROGRAM FLOW FOR PROPOSED SYSTEM

The program flow of the proposed system for identifying compromised URLs in electronic mails using Bayesian classification is analyzed below:

1. **Enter email sample** in [http://tosin.gidibusinesspages.com.ng/]: An e-mail that appears suspicious is copied by the user and pasted into the proposed system.
2. **Check URL:**Checks through the e-mail for a link starting with /http/ or /https/ if found proceed to (3)
3. **Extract URL:**Extract every string starting from the http and stop after a space character, then proceed to (4)
4. **Validate URL using REGEX:** Use regex to validate the link and if URL is valid proceed to (5).

5. **Check URL Length:** Check if URL length exceeds 55 characters, if so echo out result and proceed to (6)
6. **Check if URL has IP:** Check if URL contains an IP address, if so echo out result then proceed to (7)
7. **Check number of dots in URL:**If number of dots is greater than 5 echo out then proceed to (8)
8. **Check Age of domain:** Check the age of a domain and proceed to (9)
9. **Pare URL into a switch statement:** Using a switch statement check if URL contains suspicious characters (@, %,.) or phishing keywords (account, confirm, secure, verify, sign in, password) and if so echo out result else echo default.

The proposed system involves three modules: data module, feature extraction module and classification module.

The data module involves the collection of URLs in the e-mail. E-mails that lack URL

information are considered legitimate. The collected URLs are then transferred to the feature extraction module.

In the feature extraction module, the URL features discussed in the previous section are extracted from the URL for classification. The extracted features are stored as input and passed to the classification module. In the classification module, unknown URL given for testing is submitted to extract features associated with URL, Bayesian classifier determines whether a URL is malicious by extracting the feature values through the predefined URL-based features. Those feature values are inputted to the classifier. The classifier determines whether a new URL is phishing based on the available information. It

then alerts the user about the classification result.

VII. RESULTS ANALYSIS

In order to test the effectiveness of the proposed system, 400 URLs from PhishTank database [17] were tested and the results were analyzed in this section. We collected phishing URLs that were submitted between June 2016 and August 2016. PhishTank which is operated by OpenDNS, is a database that records the URLs of suspected websites that have been reported, the time of that report and sometimes the screenshots of the website as shown in Figure 2.

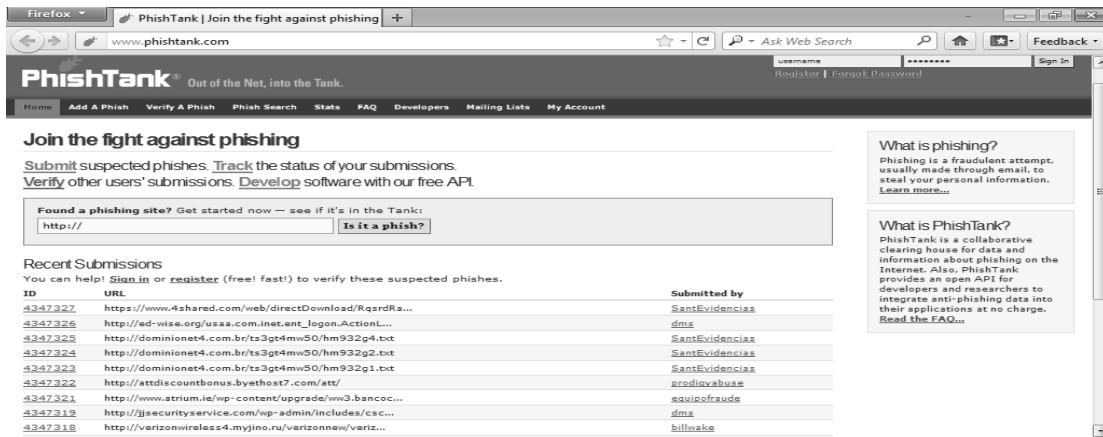


Figure 2: Screenshot of the Phishing Data Source used [17].

Table 2: Data used for the Experiments

Total Samples	405
Total Legitimate e-mails	167
Total Malicious e-mails	238
Total Training Samples	100
Total Testing Samples	305

Table 3: When the No. of tested URLs=55, Malicious URLs=34, Legitimate URLs=21

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	17
HTTP Status not ok	7
URL contains an IP address	2
Number of dots more than 5	0
Domain age not checked	5
URL contains a phishing keyword	3

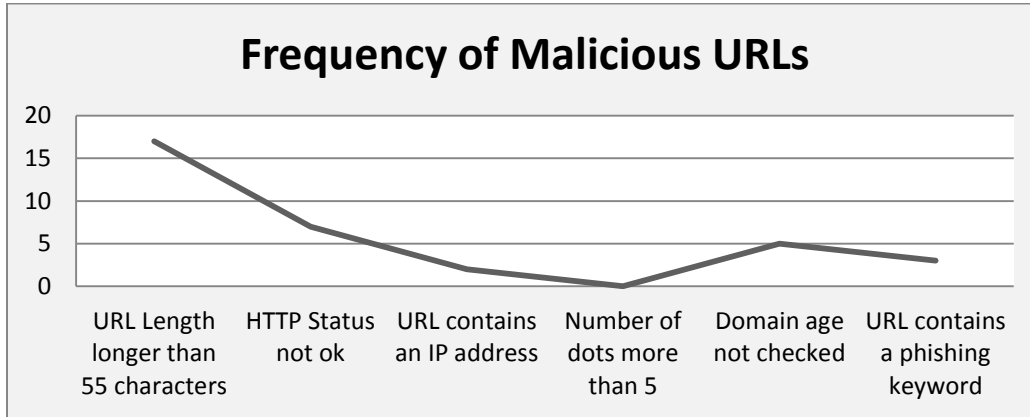


Figure 3: Graph of 55 tested URLs with 34 malicious URLs and 21 valid URLs

Table 4: When the No. of tested URLs=102, Malicious URLs=65, Legitimate URLs=37

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	21
HTTP Status not ok	8
URL contains an IP address	0
Number of dots more than 5	0
Domain age not checked	4
URL contains a phishing keyword	1

Table 3 shows that when 55 URLs were tested, the application returned 34 URLs as malicious and 21 URLs as legitimate. It can be seen from the graph of Figure 3 that the feature ‘length of URL’ occurred most in the malicious URLs followed by the HTTP status, domain age, phishing keyword and IP address features. None of the tested URLs had more than five dots.

As the number of tested URLs was increased, the application detected more malicious URLs than legitimate URLs based on the input. The graph of Figure 4 shows that the feature with the most occurrences is the URL length followed by the HTTP status, domain age and phishing keyword. None of

the tested URLs contained an IP address and the number of dots in these URLs did not exceed the given threshold.

When the number of tested URLs was increased to 152, the application detected 95 URLs as malicious and 57 as legitimate. The graph of Figure 5 shows that with the increase in the total number of tested URLs; most of the URLs were classified as malicious based on the URL length exceeding the given threshold. Other features that helped in the classification are the HTTP status and URL phishing keyword. Other detailed and similar information could be seen on Figure 6 to Figure 10 with their corresponding plotted values on Table 6 to Table 10 respectively.

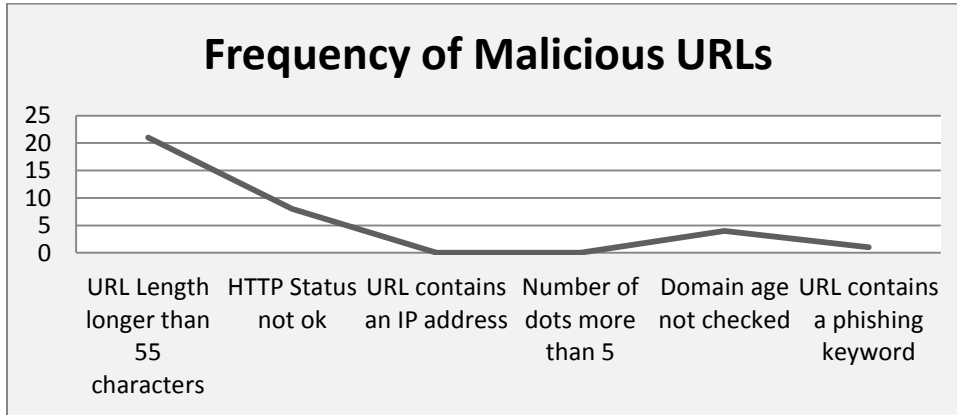


Figure 4: Graph of 102 tested URLs with 65 malicious URLs and 37 valid URLs

Table 5: When the No. of tested URLs=152, Malicious URLs=95, Legitimate URLs=57

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	21
HTTP Status not ok	8
URL contains an IP address	0
Number of dots more than 5	0
Domain age not checked	0
URL contains a phishing keyword	2

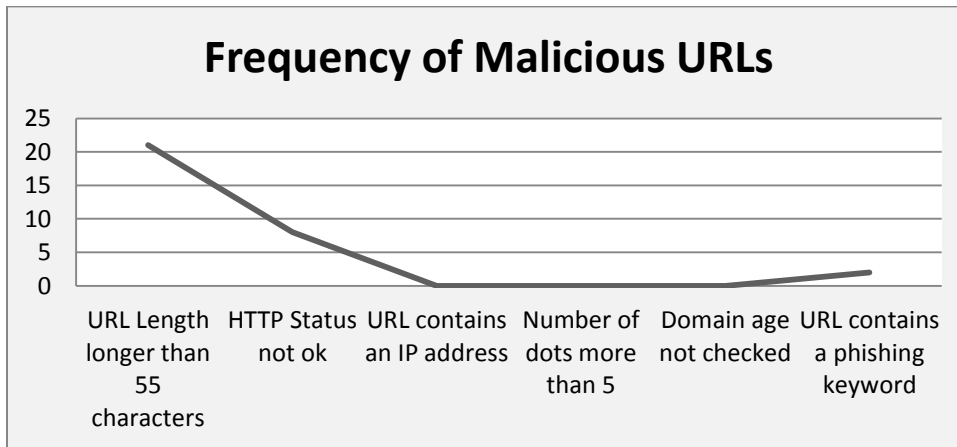


Figure 5: Graph of 152 tested URLs with 95 malicious URLs and 57 valid URLs

Table 6: When the No. of tested URLs=205, Malicious URLs=118, Legitimate URLs=87

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	17
HTTP Status not ok	4
URL contains an IP address	0
Number of dots more than 5	0
Domain age not checked	2
URL contains a phishing keyword	1

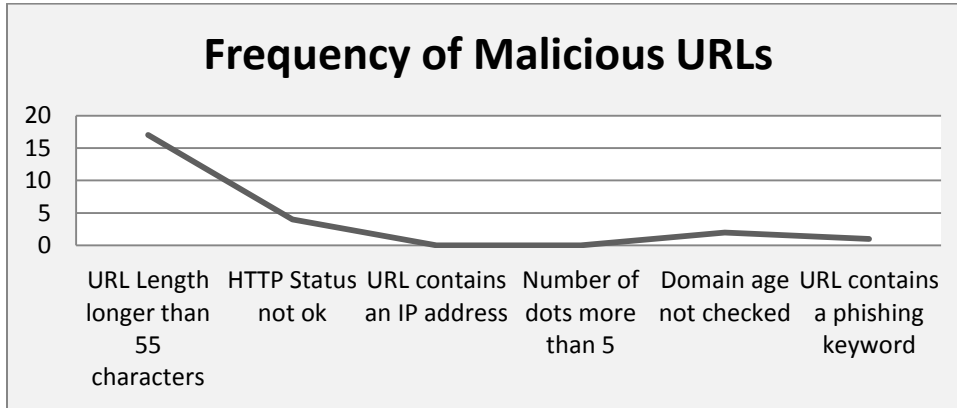


Figure 6: Graph of 205 tested URLs with 118 malicious URLs and 87 valid URLs

Table 7: When the No. of tested URLs=255, Malicious URLs=148, Legitimate URLs=107

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	15
HTTP Status not ok	8
URL contains an IP address	1
Number of dots more than 5	0
Domain age not checked	5
URL contains a phishing keyword	1

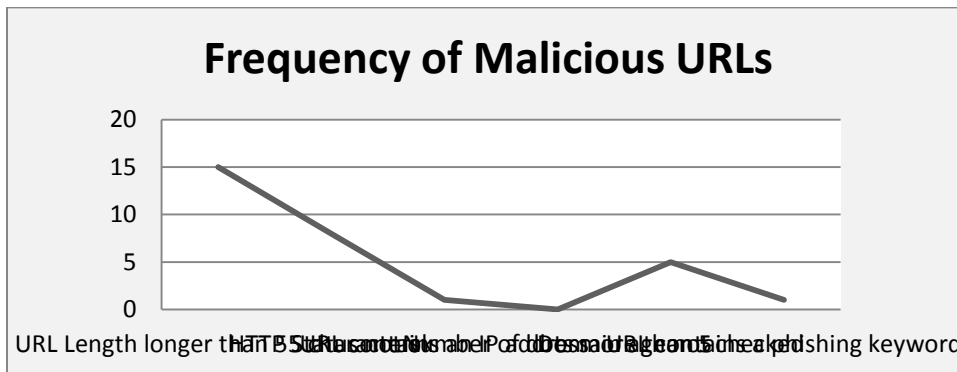


Figure 7: Graph of 255 tested URLs with 148 malicious URLs and 107 valid URLs

Table 8: When the No. of tested URLs=305, Malicious URLs=179, Legitimate URLs=126

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	26
HTTP Status not ok	4
URL contains an IP address	2
Number of dots more than 5	0
Domain age not checked	6
URL contains a phishing keyword	1

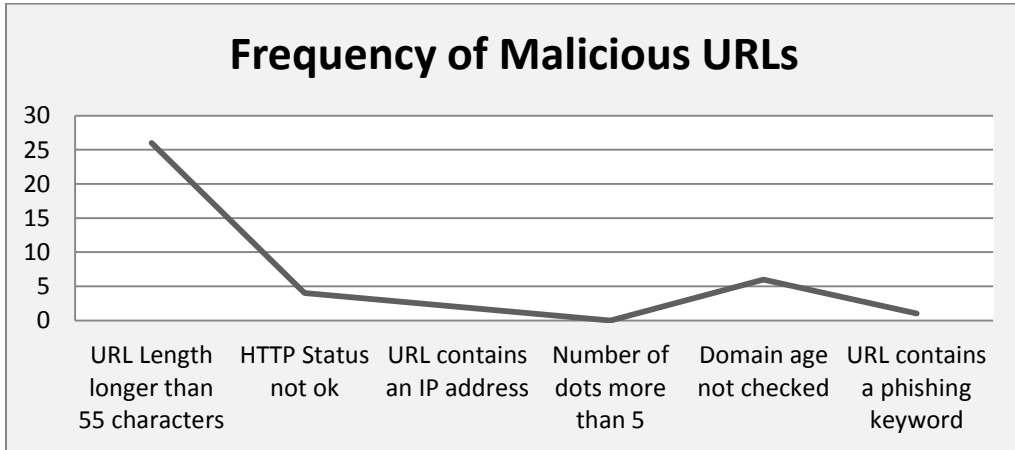


Figure 8: Graph of 305 tested URLs with 179 malicious URLs and 126 valid URLs

Table 9: When the No. of tested URLs=355, Malicious URLs=214, Legitimate URLs=141

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	26
HTTP Status not ok	7
URL contains an IP address	2
Number of dots more than 5	0
Domain age not checked	1
URL contains a phishing keyword	1

As the number of tested URLs was increased, more malicious URLs were detected than legitimate URLs. From the graph of Figure 6, it can be seen that the feature with the most occurrences is the URL length followed by the HTTP status, domain age and phishing keyword. None of the tested URLs contained an IP address and the number of dots in these URLs did not exceed the given threshold.

With more URLs being tested, there was more increase in the number of malicious URLs detected by the application than legitimate URLs. The graph of Figure 9 shows that when the total number of tested URLs was 255, the application detected 148 URLs as malicious and 107 as legitimate. Also, the malicious URLs were mostly classified based on their length exceeding the set threshold.

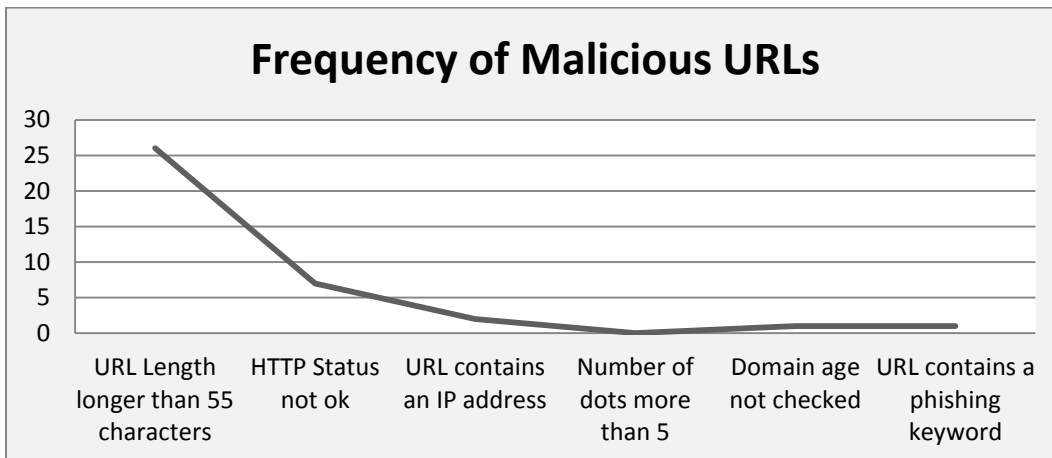


Figure 9: Graph of 355 tested URLs with 214 malicious URLs and 141 valid URLs

Table 10: When the No. of tested URLs=405, Malicious URLs=238, Legitimate URLs=167

Features	Frequency of Malicious URLs
URL Length longer than 55 characters	15
HTTP Status not ok	3
URL contains an IP address	0
Number of dots more than 5	0
Domain age not checked	1
URL contains a phishing keyword	1

Table 11: Summary of experiments carried out for F1-F6

FEATURES	R1	R2	R3	R4	R5	R5	R6	R7	R8
F1	17	21	21	17	15	15	26	26	15
F2	7	8	8	4	8	8	4	7	3
F3	2	0	0	0	1	1	2	2	0
F4	0	0	0	0	0	0	0	0	0
F5	5	4	0	2	5	5	6	1	1
F6	3	1	2	1	1	1	1	1	1

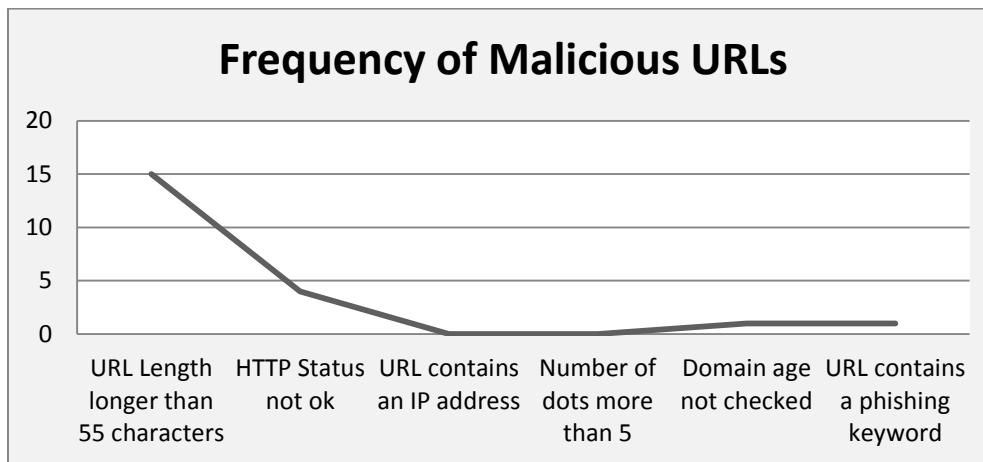


Figure 10: Graph of 405 tested URLs with 238 malicious URLs and 167 valid URLs

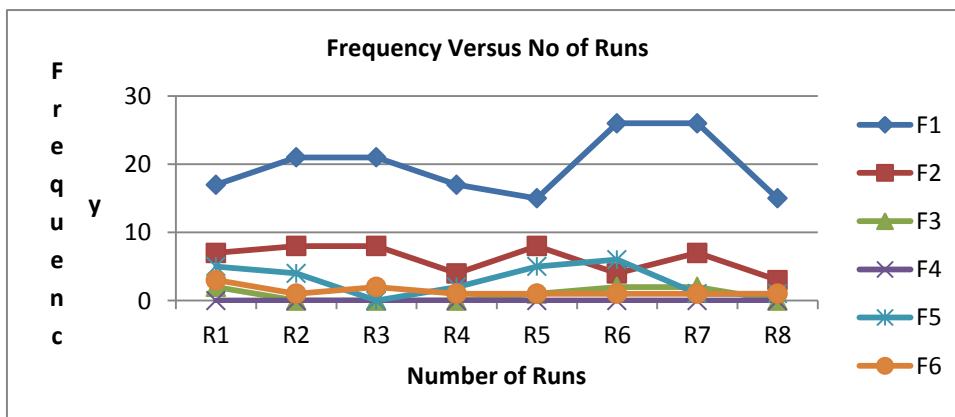


Figure 11: Graph showing the summary of experiments carried out for F1-F6

Altogether, a total of eight (8) runs were carried out. The details of the runs with the frequencies obtained for each of the features (F1-F6) are summarized in Table 11. The graphical representation is available in Figure 11.

VIII. CONCLUSION

This paper proposed a system for identifying compromised URLs in electronic mails using Naïve Bayesian Classifier, and it yielded a probabilistic phishing detection framework that can quickly adapt to new attacks with reasonably good true positive rates and close to zero false positive rates. The system

exploits the lexical and host-based features of URLs. Additionally, several machine learning algorithms were reviewed and the Naïve Bayesian Classifier was chosen to be the best for this research. The Naïve Bayesian Classifier was applied to classify suspected URLs as valid or malicious.

IX. ACKNOWLEDGMENT

The authors wish to acknowledge the financial support provided for this research through the research grant No CRC/TETFUND/NO. 2016/01 approved by the Academic Planning Unit of the University of Lagos, Lagos, Nigeria.

X. REFERENCES

- [1] Alkhozai, M. G., & Batarfi, O. A. (2011). Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code. *International Journal of Information and Communication Technology Research* , 1 (6), 283-291.
- [2] Aburrou, M., Hossain, M., Keshav, D., & Fadi, T. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert Systems with Applications* (37), 7913–7921.
- [3] Anti-Phishing Working Group. (2015). APWG Phishing Activity Trends Report - APWG.org. Retrieved from https://docs.apwg.org/reports/apwg_trends_report_q4_2015.pdf
- [4] Badra, M., El-Sawda, S., & Hajjeh, I. (2007). Phishing attacks and solutions. In *Proceedings of the 3rd International Conference on Mobile Multimedia Communications* (p. 42). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
- [5] Dhanalakshmi, R., & Chellappan, C. (2013). Detecting Malicious URLs in E-mail - An Implementation. *AASRI Procedia* , 125-131.
- [6] Dongsong , Z., Zhijun , Y., Hansi, J., & Taeha, K. (2014). A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. *Information & Management* (51), 845–853.
- [7] Gowtham, R., & Krishnamurthi, I. (2014). A comprehensive and efficacious architecture for detecting phishing webpages. *Computers & Security* , 23-37.
- [8] Han, W., Cao, Y., Bertino, E., & Yong, J. (2012). Using automated individual white-list to protect web digital identities. *Expert Systems with Applications* , 11861–11869.
- [9] Indiana University . (2014, 12 18). What is a URL? Retrieved May 12, 2016, from Indiana University Knowledge Base: <https://kb.iu.edu/d/adnz>
- [10] Jiawei Han, Micheline Kamber. (2006). *Data Mining: Concepts and Techniques*. San Francisco: Diane Cerra.
- [11] Jin-Lee, L., Dong-Hyun, K., & Chang-Hoon, L. (2015). Heuristic-based

- Approach for Phishing Site Detection Using URL Features. Proceedings of the Third International Conference on Advances in Computing, Electronics and Electrical Technology - CEET 2015 , 131-135.
- [12] Kausar, F., Al-Otaibi, B., Al-Qadi, A., & Al-Dossari, N. (2014). Hybrid Client Side Phishing Websites Detection Approach. *International Journal of Advanced Computer Science and Applications (IJACSA)* , 5 (7), 132-140.
- [13] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen, & Minh Hoang Nguyen. (2014). A novel approach for phishing detection using URL-based heuristic. *Computing, Management and Telecommunications (ComManTel), 2014 International Conference on IEEE .*
- [14] Martin, A., Anutthamaa, N., Sathyavathy, M., Marie Manjari, S., & Venkatesa, P. (2011). A Framework for Predicting Phishing Websites Using Neural Networks. *International Journal of Computer Science Issues (IJCSI)* , 8 (2), 330-336.
- [15] Narendra, M. S., Chaitali , S., Mrunal , M., & Shruti , R. (2015). AN IDEAL APPROACH FOR DETECTION AND PREVENTION OF PHISHING ATTACKS. *Procedia Computer Science* (49), 82 – 91.
- [16] Neda, A., Aladdin, A., & Fadi, T. (2014). Phishing detection based Associative Classification data mining. *Expert Systems with Applications* (41), 5948–5959.
- [17] PhishTank. (n.d.). PhishTank. Retrieved July 2016, from <http://www.phishtank.com>
- [18] Ram, B., Srinivas, M., & Andrew , H. S. (2008). Detection of Phishing Attacks: A Machine Learning Approach. *Soft Computing Applications in Industry* , 373–383.
- [19] Ramesh, G., Krishnamurthi, I., & Kumar, S. K. (2014). An efficacious method for detecting phishing webpages through target domain identification. *Decision Support Systems* , 12-22.
- [20] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian Approach to Filtering Junk E-mail. *AAAI Workshop on Learning for Text Categorization*. Madison, Wisconsin.
- [21] Sheng, S., Wardman, B., Warner, G., Cranor, L., Hong, J., & Zhang, C. (2009). An empirical analysis of phishing blacklists. In *CEAS*.
- [22] Xiang, G., & Hong, J. (2009). A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval. *Madrid, Spain: International World Wide Web Conference Committee (IW3C2)*.
- [23] Yaser, A. N. (2013). An Introduction to Electronic Commerce. *International Journal of Scientific & Technology Research* , 2 (4), 190-193.
- [24] Zhang, Y., Hong, J., & Cranor, L. (2007). CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. *Banff, Alberta, Canada: International World Wide Web Conference Committee (IW3C2)*.

AANtID: AN ALTERNATIVE APPROACH TO NETWORK INTRUSION DETECTION

N.A Azeez¹ and A.B Babatope²

^{1,2}*Department of Computer Sciences, University of Lagos, Lagos, Nigeria*

¹*nazez@unilag.edu.ng; ²dudeolu@gmail.com*

ABSTRACT

Information that is not properly secured has the tendency of being vulnerable to intrusions and threats. Security has become not just a feature of an information system, but the core and a necessity especially the systems that communicate and transmit data over the Internet for they are more susceptible to intrusions and threats. This work aims at presenting an approach to intrusion detection. This paper presents an Intrusion Detection System (IDS) using Genetic Algorithm (GA). GA was chosen because it has been proven to efficiently detect different types of intrusions. GA parameters and the evolution process are discussed in detail. The DARPA 1998 dataset was used to implement and measure the performance of the system. The result that was obtained showed that specific Class of IP addresses were more susceptible to intrusions and threats.

Keywords: *Intrusion, Genetic Algorithm, detection, Security, DARPA dataset*

1.0 INTRODUCTION

In the world today, information is fundamental for basic operations in every institution, organization and the society at large. Information used to be stored in traditional paper file format. This method did not help make business operations to be effective and efficient. In recent times, Information involves computers, networks and communication media which are used to transmit the data from one point to another. Information systems are used to store the data while the network is used for communication among information systems. Organizations are becoming more dependent on information systems and computer networks for the storage and transmission of information which has increased the risk involved with the use of information systems and computer networks. The advancement of computer network technology has caused it to be a target to attacks from unauthorized entities.

The issue of information and network security has been a major concern for the information technology society and has

prompted stakeholders to strengthen their defense against unauthorized entities such as hackers, Trojan horses, malicious programs, viruses etc. These unauthorized entities, after gaining access to computer networks, have caused serious problems and cost the stakeholders resources in form of time and money. With the recount of past intrusion events, it is better to take proactive steps to prevent malicious and unauthorized entities access to the network. The most popular technique that is used to guard against unauthorized entities is the use of firewall [2]. The firewall blocks strange users from gaining access to the network. In recent years, it has been discovered that the firewall is not enough to keep out unauthorized entities from a computer network. Intrusion detection system (IDS) is one of the new techniques used to secure network systems [2]. The IDSs is an efficient method of securing computer networks. It observes events happening within the network and takes record of information relating to each event that happens within the network. It reviews the records to look out for unusual

events so it can notify the network security administrators of such events through the reports it produces [5].

2.0 LITERATURE REVIEW

According to Scarfone and Mell (2007) “Intrusion detection is the process of monitoring the events occurring in a

computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies or standard security practices” and the Intrusion Detection System (IDS) is the software that automates the intrusion detection process [12].

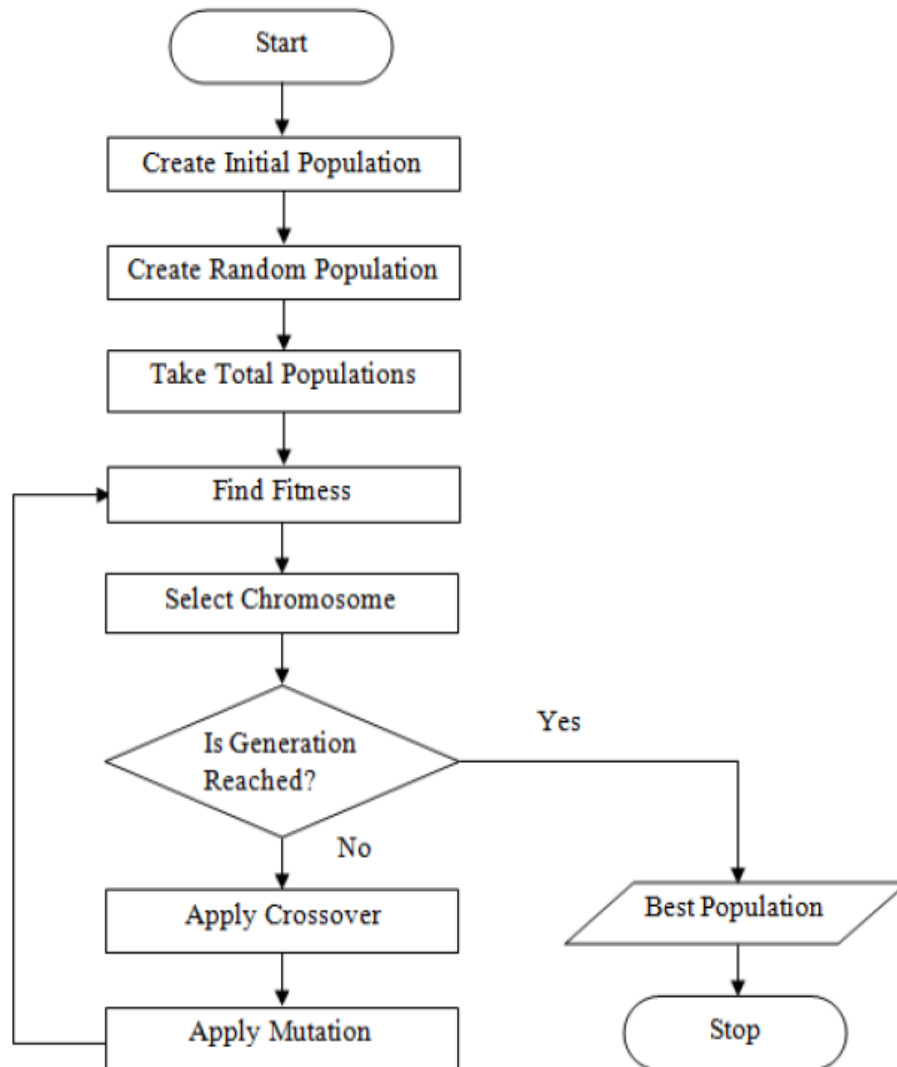


Figure 1: Flowchart of the Genetic Algorithm (Hassan, 2013)

According to Rajan & Cherukuri (2010), the commercial Intrusion Detection Systems usually merge the two approaches in one system just because the real time intrusion detection involves a larger scope. This approach involved the use of evolution theory and information evolution to screen the traffic data in order to reduce the

complexity of the network.

In a research work, Hassan (2013) developed an Intrusion Detection System where genetic algorithm and fuzzy logic were applied. The proposed Genetic algorithm [6] for this system comprised of two modules that each served at different stages; the training stage and detection stage

[4]. In the training stage, the classification rules were developed from the network audit data and in the detection stage, the rules which have been generated were used to classify the incoming network in real-time [5].

In the research work done by Sharma and Gupta (2012), they looked at using the artificial intelligence approach for intrusion detection in soft systems. Artificial intelligence has been used to solve complex problems such as intrusion detection [7] just because of its ability to learn and evolve as it gets more information [13]. The type of threat that was dealt with in this paper is Black holes. Black holes refer the nodes in a network that disturb the transmission of data by capturing the data.

Evaluation of the system was measured using the following metrics

$$\begin{aligned} & \text{Transmissiondelays} \\ & = \text{timeofpathhalt} \\ & * \text{numberofnodes} \end{aligned}$$

$$\begin{aligned} & \text{Intrusionrate} \\ & = \beta * \text{delays} \\ & * \text{transmissiondelays} \end{aligned}$$

Where,
delays

$$= \frac{1}{(\text{linkspeed})(N_p - N_t)N + (N - 1)D_l} \dots\dots\dots 1$$

N is the number of nodes,
 N_t is the number of retransmissions,
 N_p is the packet size and
 D_l is the average delay that is measured taking into account the ideal conditions

The result obtained is unreliable as there was a contradiction and inconsistencies in the final output.

Li (2004) also reported an intrusion detection system using genetic algorithm to detect anomalous network intrusion. The method used for the system includes quantitative and categorical features of network data for obtaining classification rules. The use of quantitative feature can improve the detection rate of the system although experimental results to support the statement are not available. Unlike the other implementations that have been discussed,

this implementation involved both temporal and spatial information of network connections in encoding the network connection information into rules in Intrusion Detection System. This was of aid to the detection of complicated network anomalous behaviors. This work as limited area of application as it focused on the TCP/IP network protocols [8].

Lu & Traore (2004) proposed a rule evolution approach based on Genetic Programming (GP) for detecting intrusions on computer networks. Four genetic operators were used to evolve the new rules; reproduction, mutation, crossover and dropping condition operators. They went further using support-confidence framework as the fitness function and at the end, the classification of connections was accurate. The DAPRA dataset was used as well to implement and evaluate the system [9]. The main drawback of this research is the random selection of crossover and mutation points in the system run, thereby reducing the detection rate of the system [11].

The paper by Abdullah et al. (2009) described the use of Genetic algorithm to derive classification rules for intrusion detection using information theory to filter the network traffic data. The system was in two phases; the pre-processing and features extraction phase and the training and testing phase. Unlike the proposed system, this system used KDD99 benchmark dataset for the evaluation of the system and the result gotten at the end of the experiment showed that the detection rate was up to 99.87% and the false positive rate at 0.003% [1].

The experiment carried out by Chittur (2001) was to analyse the effectiveness of using Genetic algorithm for computer network intrusion detection system. Genetic algorithm has become one of the novel approaches to solving the intrusion detection problem as it was also used in this system. The system created an empirical model of malicious programs/ behavior using the training data (i.e. 1999 Knowledge Discovery in Database KDD cup data); a process of training the system in order to

generate rules that will be used during the testing process to recognize intrusive activities which have not been seen by the system before. The result at the end of the

experiment showed an overall accuracy level of 97.8%; high detection rate and low false positive rate.

Algorithm I: Tournament selection

Input: Population of chromosome

Output: selected chromosome for crossover

- i. Select 3-chromosomes from the population at random*
 - ii. Select the best 2-chromosome based on the fitness function value*
 - iii. Return the selected two chromosomes*
 - iv. Apply Crossover / Select best chromosome to be parent*
-

2.1 SELECTION (or Reproduction)

For the proposed system, the tournament selection was adopted. In this selection method, the chromosomes are selected randomly from the present generation of individuals so that a lesser number of chromosomes are selected in the next iteration. The tournament selection method is implemented as follows [10].

2.2 CROSSOVER (or Recombination)

This operator is used to combine two chromosomes which have been selected using the initial operator for them to produce a new chromosome. The new chromosomes produced, can always be better than the parent chromosomes if the best characteristics of the parent chromosomes are selected and inherited by the offspring. There are different types of Crossover methods such as; one-point, two point,

uniform, arithmetic and Heuristic crossovers. For this study, we will look at the two point crossover method.

2.3 MUTATION

For the population of each generation, there is a possibility for a change in the gene of the chromosomes. This change depends on the mutation rate of the system. The mutation operator is used to introduce genetic diversity to each generation for the purpose of distinction from one generation to the next generation. Mutation causes new gene values to be added to the gene pool, thereby causing a variation in the genes of the chromosomes. There are various mutation types such as; Flip bit, Boundary, non-uniform and Gaussian mutation methods. Mutation is implemented as follows [10].

Algorithm II: Mutation of rules

Input: A chromosome rule

Output: Same or Mutated chromosomes,..., a fns of mutation rate

- i. Set mutation threshold (between 0 and 1)*
 - ii. For each network attribute in chromosome*
 - iii. Generate a random number between 0 and 1*
 - iv. If random number > mutation threshold then*
 - v. Generate random value w.r.t data properties*
 - vi. Set chromosome attribute value with generated attribute value*
 - vii. End if*
 - viii. End for Each*
-

2.4 RULE DEFINITION

In the dataset, there are 7 attributes represented as columns. A record represents

a chromosome and each is a rule in itself. The rule is denoted in the *if(condition)then (outcome)* format. The

first 6 attributes which are Duration, protocol, Source port, Destination port, Source IP and Destination IP form the condition part of the rule. Then the last attribute; Attack name represents the

outcome part of the rule. Therefore, if a connection matches the conditions of any of the rules in the dataset, the connection is classified as the same category as the rule it matches.

Algorithm III: Genetic Algorithm classification rule

- i. *Random generation of initial chromosomes*
 - ii. *Set $w1 = 0.2$, $w2 = 0.8$, $T = 0.5$, Max Generations = 100*
 - iii. *Set $N =$ total number of record in training set*
 - iv. *Set generation counter = 0*
 - v. *For each chromosome in population*
 - vi. *Set $A = 0$, $AB = 0$*
 - vii. *For each record in dataset set*
 - viii. *If record matches chromosome*
 - ix. *$AB = AB + 1$*
 - x. *End If*
 - xi. *If record matches only condition part*
 - xii. *$A = A + 1$*
 - xiii. *End If*
 - xiv. *End for Each record*
 - xv. *End for Each chromosome*
 - xvi. *$Fitness = W1 * AB/N + W2 * AB/A$*
 - xvii. *If $Fitness > T$*
 - xviii. *Select fitted chromosomes into new selection pool*
 - xix. *End if*
 - xx. *For each chromosome in new pool/population*
 - xxi. *Select chromosome for breeding*
 - xxii. *Apply crossover and mutation to new offspring*
 - xxiii. *Place newly created chromosome into population*
 - xxiv. *End for Each*
 - xxv. *Kill old pool, new pool now current pool*
 - xxvi. *Increment generation Counter by 1*
 - xxvii. *If generation Counter < Max Generation then*
 - xxviii. *Goto line v*
-

2.5 CHROMOSOME REPRESENTATION

The chromosome is a string that is used to represent an individual in Genetic algorithm and it symbolizes a solution to the problem. A chromosome comprises of at least a gene which represents the attribute of the chromosome. Previous articles stated that the implementation of Genetic algorithm was achieved with the use of chromosome-like data structure which has also been used. T represented is shown in Table 1 where each chromosome comprises of at least a gene. The attribute 'Duration' is an example of a

chromosome with more than one gene due to the format of the attribute.

In this system, the wildcard character was used. The wildcard character is a special character that can represent a number of characters. The wildcard character representation of a gene is a way to make the rules more general thereby making it possible to use a single character to represent a number of values. The gene with a wildcard character is encoded as -1 . For example, a Source IP can have the value $\{192.1.-1.25\}$. The character -1 in the IP address represents any number between 0

and 255.

2.6 FITNESS FUNCTION

As earlier stated, the fitness function is used to calculate the fitness of the chromosome in

order to determine the chromosomes to be selected for reproduction. The fitness models that was used for this system is the support and confidence model.

Table 1: Chromosome representation

Attribute Name	Number of Genes	Format
Duration	3	H:M:S
Protocol	1	Numeric
Source Port	1	Numeric
Destination Port	1	Numeric
Source IP	4	a.b.c.d
Destination IP	4	a.b.c.d
Attack Name	1	String

The model is implemented as follows:

Algorithm IV: Fitness function algorithm

-
- i. If A then B,*
 - ii. support = $|A \text{ and } B|/N$*
 - iii. confidence = $|A \text{ and } B|/|A|$*
 - iv. fitness = $w1 * \text{support} + w2 * \text{confidence}$*
- N = Number of connections in training data*
|A| = Number of connections matching condition A
|A and B| = Connections matching rule if A and B
w1, w2 = Weights to balance/control the two terms
-

3.0 IMPLEMENTATION: How GA was linked with Intrusion Detection

The GA was linked with intrusion detection by using the GA algorithm to classify the different network connections that the system comes across. Therefore, the GA makes it possible for the intrusion detection system to differentiate the different types of network connections. Chromosome spans feature space as each chromosome represents a single network connection which also has different attributes like source port, source IP address, Protocol etc. Therefore, each chromosome can be equated to a network connection. Fitness function is used to determine the mostfit set of chromosomes in respect to other chromosomes present, that will be used for

recombination in the next generation.

4.0 DARPA DATASET

The DARPA dataset was created in 1998 out of the need to evaluate intrusion detection systems by the Lincoln Laboratory of MIT and was first made available to the general public in February 1998 [14]. These data were created by connecting host systems with a traffic generator in order to simulate a small US Air force base of limited personnel, connected to the Internet [3]. The dataset is made up of 7 columns namely; Duration, Protocol, Source port, Destination port, Source IP, Destination IP and attack name. Duration is the period of time it takes for data to be transmitted from source system to destination system. The duration can be recorded in seconds, minutes and

hours

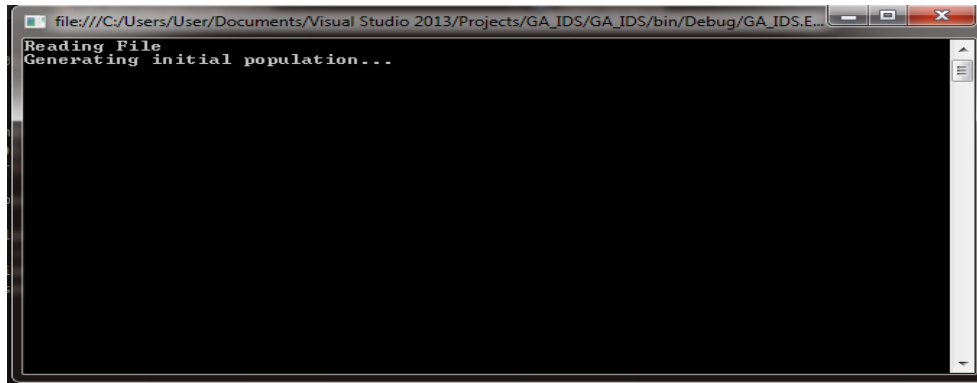


Figure 2: Initialization of the program

As seen in Figure 2, on initialization, the system randomly generates a population of individuals and reads the data from the dataset file. The randomly generated individuals represent the random connection records that will be subjected to the Intrusion Detection system so the records can be checked for intrusions. On the other hand, the dataset will serve as the benchmark for comparison. Each random individual will be scanned against every record in the dataset file and if there is match, the random

individual will be identified as the record it matches. Without reading the file, the program will not progress and will not show the 'reading file' message as seen in the Figure 3:

The progress of the system will show the results of applying genetic algorithm to detect intrusion. The result window; Figure 4 shows the entire gene (attributes) of each individual as well as the attack name or null if it is a normal connection.

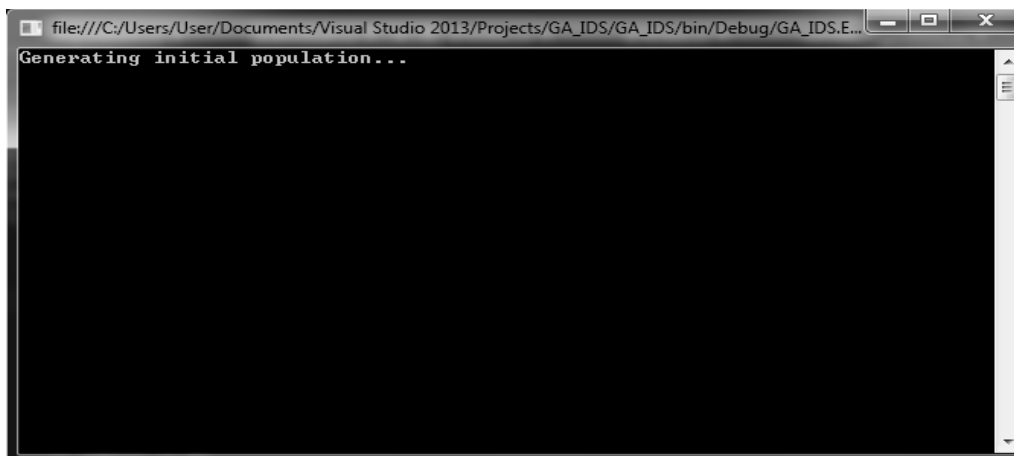


Figure 3: Initialization without reading the dataset file

The screenshot shows a log window with the following data:

Duration	Protocol	Source Port	Destination Port	Source IP	Destination IP	ATTACK
0:0:53	ftp-data	38127	1985	65.127.-1.203	243.192.194.177	rcp
-1:1:0	auth	26586	55979	20.154.30.-1	51.70.111.254	rsh
0:0:-1	rsh	62512	26370	160.72.57.42	135.-1.237.4	phf
-1:1:0	auth	26586	-1	242.41.151.123	239.18.88.223	guess
0:0:-1	rsh	62728	55979	20.154.30.-1	51.70.111.254	rsh
0:0:-1	http	-1	12106	104.65.17.124	-1.191.-1.244	port-scan
0:0:39	rsh	62728	-1	242.41.151.123	239.18.88.223	guess
0:0:39	http	-1	12106	104.65.17.124	-1.191.-1.244	port-scan
0:-1:30	telnet	12106	41499	219.158.247.45	63.46.33.220	phf
0:-1:30	telnet	12106	41499	219.158.247.45	63.46.33.220	phf
0:0:-1	rsh	62728	26586	161.188.35.228	-1.192.104.220	phf

Figure 4: The result of the detection system

The first column is the duration of the network connection which is in the H:M:S format. The second column is the protocol of the connection such as File Transfer Protocol (FTP), Simple Mail Transfer Protocol (SMTP), Remote shell (rsh). A protocol is a means for two or more entities on the network system to communicate. The third and fourth columns are the destination port and source port of the connection respectively which is used to identify a particular type of service on the network. The fifth and sixth columns are the source IP address and destination IP address. The last

column, which is the seventh column, provides the name of the different attacks which have been identified by the intrusion detection system. A brief summary of the dataset file is given in Table 2:

As previously stated that there are different categories of network attacks which are Denial of Service (DoS), Remote to User Attacks (R2L), User to Root Attacks (U2R) and Probe. The intrusions contained in the dataset are of different network attack categories which are shown in Table 3 and a graphical representation of the table Figure 5 follows immediately;

Table 2: Distribution of the network connections in the dataset

Normal	256
Intrusion	39
Total	295

Table 3: Distribution of the Intrusions

Probe	R2L	U2R	
Portscan – 30	Phf – 1 Guess – 4	rlogin – 1 rsh – 2 rcp – 1	
30	5	4	Total = 39

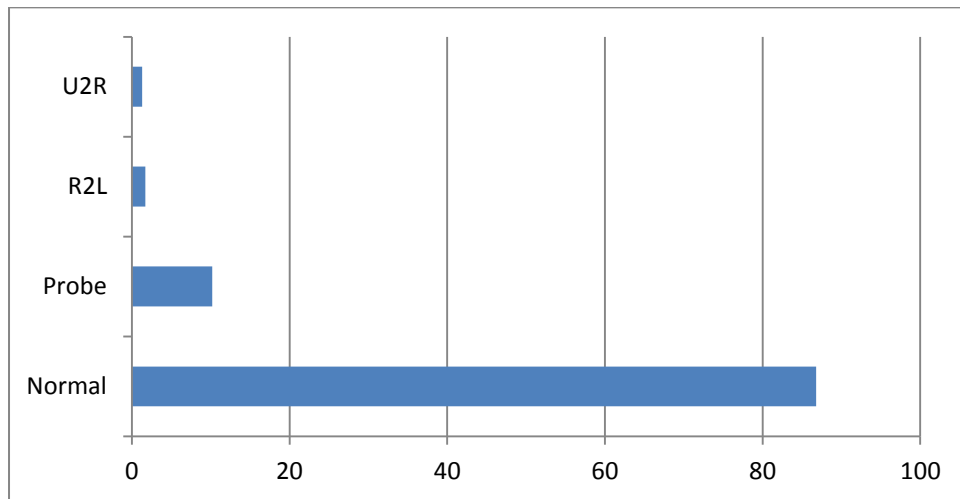


Figure 5: Graphical distribution of the network connections.

5.0 ANALYSIS OF RESULTS

The system was run a number of times and four results were selected to be analyzed for other findings and note. Each set of result consists of 50 records. For each set of results, the IP addresses of the source IP address and destination IP addresses were classified under the Standard IP address classes and a graphical representation to show the relationship between the two addresses based on the classes, and a table and bar chart to show the relationship

between protocols and the type of intrusions.

5.1 The first run

Table 4 shows classes which the source and destination IP addresses fall and Figure 6 is a graphical representation.

Table 5 shows the distribution of attacks against the Protocols and Figure 7 shows a graphical representation of the distribution.

Table 4: Classification of IP addresses of the first set results

	Source IP	Destination IP
Class A	20	50
Class B	30	0
Class C	0	0
Class D	0	0
Class E	0	0

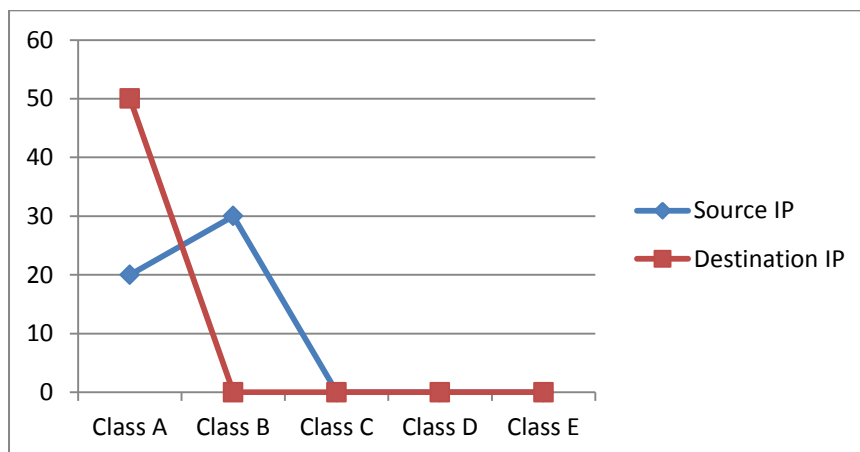


Figure 6: Graphical classification of the IP addresses of the first set of results

Table 5: Distribution of intrusion attacks to the Protocols for the first set of results

Protocols	Attack names					
	Phf	Port scan	rcp	rlogin	rsh	
ftp						
ftp-data		2	3		1	6
http		2	4		14	20
Rsh		3	2		3	8
Sntp		5	2		5	12
telnet			1		3	4
		12	12		26	

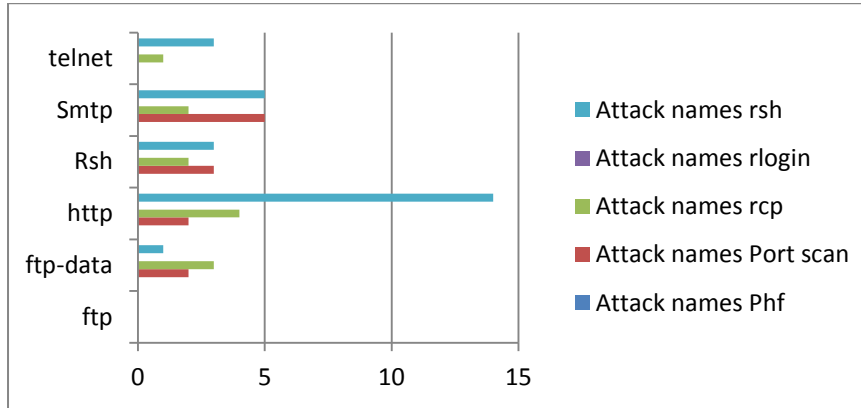


Figure 7: Graphical representation of the distribution of attacks to protocols for the first set of results

Table 6: Classification of IP addresses of the second set results

	Source IP	Destination IP
Class A	19	39
Class B	0	11
Class C	0	0
Class D	0	0
Class E	31	0

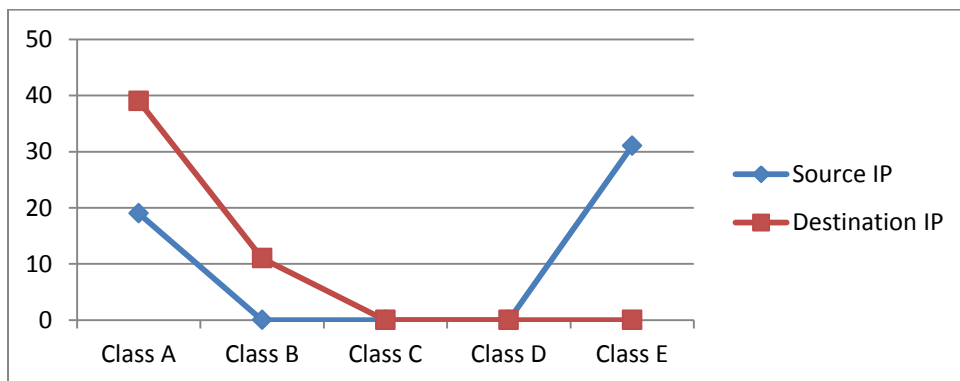


Figure 8: Graphical classification of the IP addresses of the second set of results

5.2 The second run;

Table 6 shows classes which the source and destination IP addresses fall and Figure 8 is a graphical representation of the table.

Table 7 shows the distribution of attacks against the Protocols and Figure 9 shows a graphical representation of the distribution.

Table 7: Distribution of intrusion attacks to the Protocols for the second set of results

Protocols	Attack names					
	Phf	Port scan	rcp	rlogin	rsh	
ftp		2		4		6
ftp-data		11		21		32
http						
Rsh		8		4		12
Smtpt						
telnet						
		21		29		

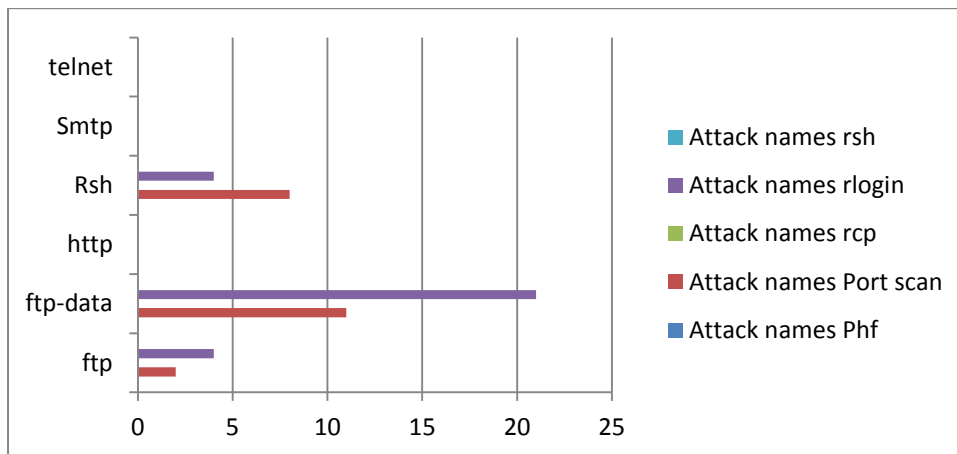


Figure 9: Graphical representation of the distribution of attacks to protocols for the second set of results

Table 8: Classification of IP addresses of the third set results

	Source IP	Destination IP
Class A	21	50
Class B	8	0
Class C	21	0
Class D	0	0
Class E	0	0

5.3 The third run

Table 8 shows classes which the source and destination IP addresses fall and Figure 10 is a graphical representation of the table.

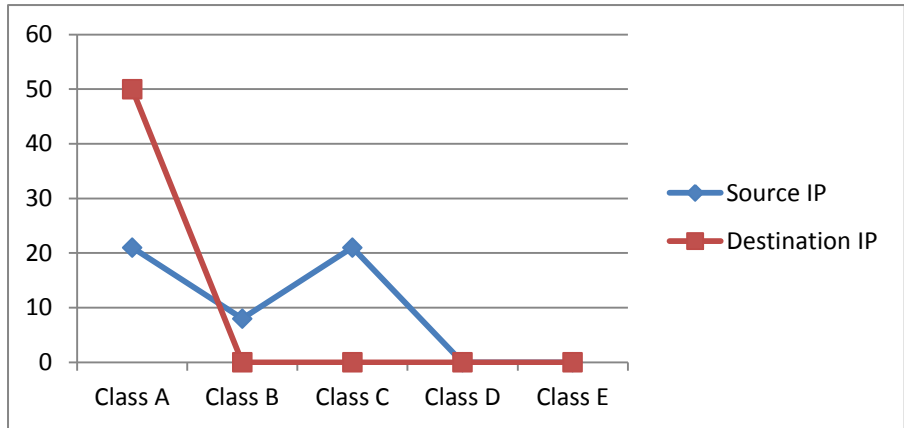


Figure 10: Graphical classification of the IP addresses of the third set of results

Table 9: Distribution of intrusion attacks to the Protocols for the third set of results

Protocols	Attack names					
	Phf	Port scan	rcp	rlogin	rsh	
ftp						
ftp-data						
http						
Rsh		1	8		16	25
Smtpt						
telnet		1	9		15	25
		2	17		31	

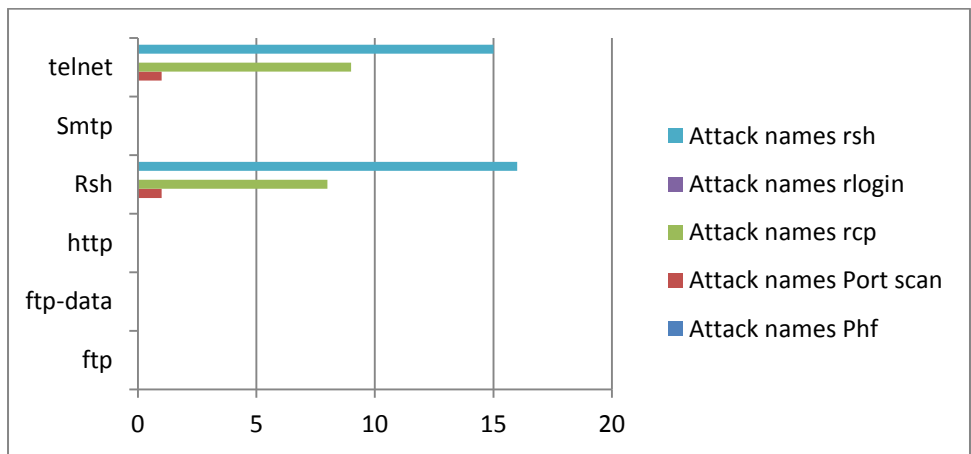


Figure 11: Graphical representation of the distribution of attacks to protocols for the third set of results

Table 9 shows the distribution of attacks against the Protocols and Figure 11 shows a graphical representation of the distribution.

5.4 Fourth run;

Table 10 shows classes which the source and destination IP addresses fall and Figure 12 is a graphical representation of the table.

Table 11 shows the distribution of attacks against the Protocols and Figure 13 shows a graphical representation of the distribution.

Table 10: Classification of IP addresses of the fourth set results

	Source IP	Destination IP
Class A	46	44
Class B	4	0
Class C	0	0
Class D	0	0
Class E	0	6

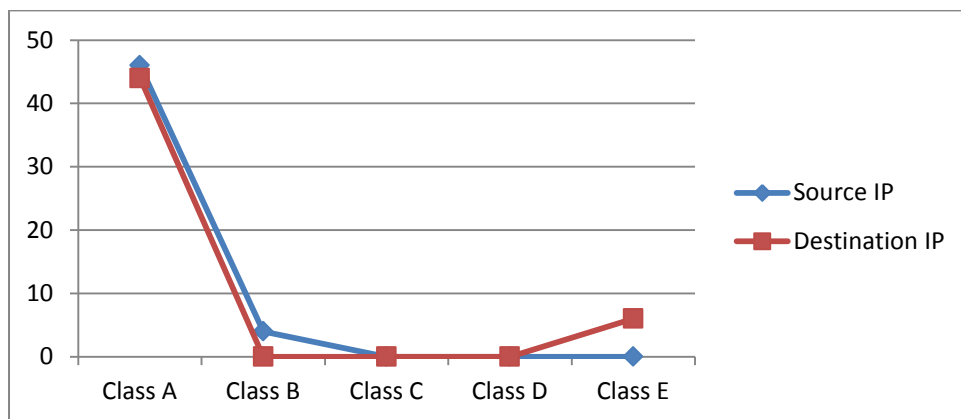


Figure 12: Graphical classification of the IP addresses of the fourth set of results

Table 11: Distribution of intrusion attacks to the Protocols for the fourth set of results

Protocols	Attack names					
	Phf	Port scan	rcp	rlogin	rsh	
ftp						
ftp-data						
http	4	5				9
Rsh	1	1				2
Smtpt	9	30				39
telnet						
	14	36				

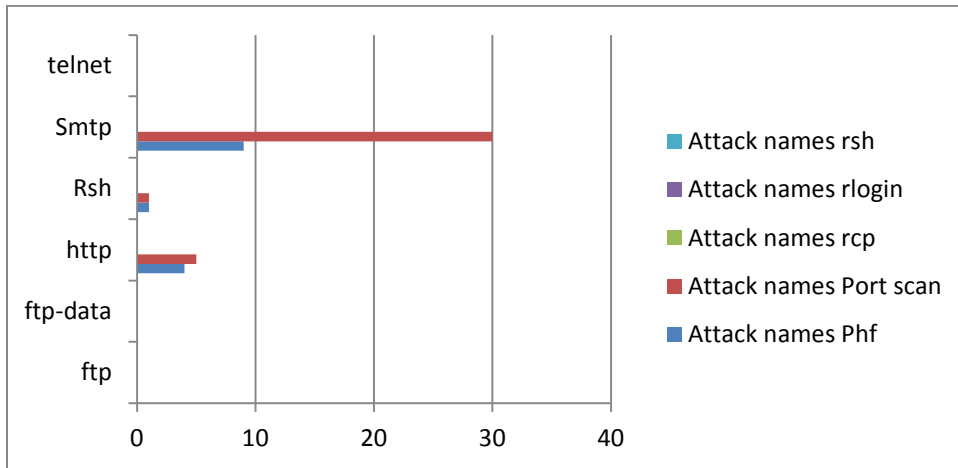


Figure 13: Graphical representation of the distribution of attacks to protocols for the second set of results

6.0 CONCLUSION

Artificial intelligence has continued to contribute immensely to technological development and the use of Genetic algorithm for intrusion detection is not exempted. It is becoming more essential to organizations for the purpose of network management in order to prevent intrusion. This is due to the fact that data is only useful when it still has its integrity and confidentiality. Therefore it is important to continue to improve on IDS.

In conclusion, from the four set of results, it is clear that most of the network activities take place mostly with the IP address class A, both the source address and destination address. Class B IP addresses were relatively busy as well based on the results. It is also worthy to note that the 'port scan' attack which falls under the Probe classification is the intrusion with the

highest frequency.

7.0 FUTURE WORK

More research on Genetic algorithm should be carried out in order to continually improve on the way classification rules are derived so as to get a better detection rate of intrusive connections in the system. More variations of genetic algorithms should be evaluated in order to compare the results from each variation and determine the best in order to have system that detects intrusions better.

8.0 ACKNOWLEDGMENT

The authors wish to acknowledge the financial support provided for this research through the research grant No CRC/TETFUND/NO. 2016/01 approved by the Academic Planning Unit of the University of Lagos, Lagos, Nigeria.

9.0 REFERENCES

- [1] Abdullah, B., Abd-alghafar, I., Salama, G. & Abd-alhafez, A., 2009. Performance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection System. *13th International Conference on Aerospace Sciences & Aviation Technology, ASAT – 13, May 26 – 28, 2009.*
- [2] Alhazzaa, L., 2007. A Literature Review on Intrusion Detection Systems using Genetic Algorithms. *CSC 590, Computer Science College, King Saud University.*
- [3] Brown, C., Cowperthwaite, A., Hijazi, A. & Somayaji, A., 2009. Analysis of the 1999 DARPA/Lincoln Laboratory IDS Evaluation Data with NetADHICT. *Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA).*
- [4] Chandrakar, O., Singh, R. & Barik, L., (2014). Application of Genetic Algorithm in Intrusion Detection System. *Control Theory and Informatics, The International Institute for Science, Technology and Education (IISTE), Vol. 4, No. 1, 2014. ISSN (online): 2225-0492*
- [5] Hassan, M.M., 2013. Network Intrusion Detection System Using Genetic Algorithm and Fuzzy Logic. *International Journal of Innovative Research in Computer and Communication Engineering (IJRCCE), Vol. 1, Issue 7, September 2013. ISSN (online): 2320-9801.*
- [6] Hoque, M.S., Mukit, A. & Bikas, A.N., 2012. An Implementation of Intrusion Detection System using Genetic Algorithm. *International Journal of Network Security & Its Applications (IJNSA), Vol. 4, No. 2, March 2012.*
- [7] Jaiganesh, V., Sumathi, P. & Vinitha, A., 2013. Classification Algorithms in Intrusion Detection System: A Survey. *International Journal of Computer Technology and Applications, Vol. 4 (5), 746-750. ISSN: 2229-6093.*
- [8] Li, W., 2004. Using Genetic Algorithm for Network Intrusion Detection. *Mississippi State University, Mississippi State, MS 39762.*
- [9] Lu, W. & Traore, I., 2004. Detecting New Forms of Network Intrusion Using Genetic Programming. *Computational Intelligence, vol. 20, pp. 3, Blackwell Publishing, Malden, pp. 475-494.*
- [10] Ojugo, A., Eboka, A., Okonta, O., Yoro, R. & Aghware F., 2012. Genetic Algorithm Rule-Based Intrusion Detection System (GAIDS). *Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 8 August. ISSN 2079-8407*
- [11] Paliwal, S. & Gupta, R., 2012. Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm. *International Journal of Computer Applications (0975-8887), Volume 60, No. 19, December 2012.*
- [12] Scarfone, K. & Mell, P., 2007. Guide to Intrusion Detection and Prevention Systems (IDPS). *National Institute of Standards and Technology NIST special publication 800-94.*
- [13] Sharma, V. & Gupta, T., 2012. An Artificial Intelligence Approach towards Intrusion Detection in Soft Systems. *International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 2, February 2012. ISSN: 2277 128X.*
- [14] Thomas, C., Sharma, V. & Balakrishnan, N., 2008. Usefulness of DARPA Dataset for Intrusion Detection System Evaluation.