2019 PROCEEDINGS

13th INTERNATIONAL MULTI-CONFERENCE ON ICT APPLICATIONS

AICTTRA 2019

17TH -20TH
NOVEMBER, 2019

Main Auditorium,
African Centre of Excellence (OAK-Park),
Obafemi Awolowo University, Ile-Ife, Nigeria

THEME:

Application of Information and Communication Technologies to Teaching, Research and Administration

FOR LEARNING AND CULTURE

ISSN 2141 - 0240

The 13th International Multi-Conference on ICT Applications

Theme:
Application of Information and Communications Technology To Teaching, Research and Administration (AICTTRA)

November 17th – 20th, 2019

@

Main Auditorium
African Centre of Excellence (OAK-Park)
Obafemi Awolowo University, Ile-Ife, Nigeria

PROCEEDINGS - Volume XII

Edited by
Professor E.R. Adagunodo
Professor G.A. Aderounmu
Professor A.I. Oluwaranti
Dr. E.A. Olajubu
Dr. B.I. Akhigbe
Dr. I.P. Gambo

Type Setting & Formatting
Mr. H.O. Ola

Organized by
Department of Computer Science & Engineering
In Collaboration with
African Centre of Excellence
- Obafemi Awolowo University -

FOREWARD

It is with great pleasure and delight that I welcome all of us to the 13th edition of the International Conference on Application of Information and Communication Technologies to Reaching, Research and Administration, tagged AICTTRA 2019, at the African Centre of Excellence (OAK-Park) Main Auditorium, Obafemi Awolowo University, Ile-Ife, Nigeria between November 17th and 20th, 2019. I understand that conferees came from different places within and outside Nigeria to attend this great event. I usually refer to the conference as the pilgrimage for ICT professionals and enthusiasts. This year’s proceedings contains articles that have been subjected to reasonable reviews to make sure that written and submitted papers for the conference are up to presentable standards.

The Programme of the conference is varied with a wide range of applications to which ICT can be put to use. This has consistently being the case since the inception of the conference, and thus endorses the fact that it is impossible to place any specific bound or limit to what ICT can be applied to. The field of ICT has continued to grow like wild fire in the Harmattan with evidences that are mind blowing in terms of its area of application. The exploration of ICT in these areas has been the main business of all the previous editions of the conference as well as this current one.

The programme of the conference has been threaded into Technical and Plenary Sessions. The sessions as usual have huge promises in that fruitful debates are stimulated on emerging and already consolidated areas of research in the use of ICT. Crosscutting issues are also included in organized syndication with informal sessions on discussions during the conference. The immediate past President of the Nigeria Computer Society (NCS), Professor G.A. Aderounmu who happen to be one of the initiators of the conference and the current Dean of the Faculty of Technology, Obafemi Awolowo University will be on hand to share his wealth of experiences in teaching, research and administration using ICT. There will be lead paper presentations by eminent researchers and practitioners in ICT. This year a Cyber Threat and Vulnerability expert from the UK, Mr. Bimbo Adebayo will be one of the Lead Speakers, and other key players in the ICT industry will also be on hand to share their ICT experiences from industry perspective.

The organizers of the event owe special thanks to the Vice-Chancellor of Obafemi Awolowo University, IIe-Ife, Nigeria, Professor E.O. Ogunbodede for his continuous support in the organization of the conference. The fathers of the conference - Professor L.O. Kehinde and Professor E.R. Adagunodo are also well acknowledged for being visionary and for their continuous support. The Local Organizing Committee members – Prof. A.I. Oluwaranti (Chairman) and members – Dr. E.A. Olajubu, Dr. I.P. Gambo, Dr. B.I. Akhigbe, Dr. B.O. Akinyemi, Dr. R.N. Ikono, Engr. Tope Ajayi, Miss A.R. Lawal, Dr. S. Aina, Dr. H.O. Odukoya are well acknowledged and appreciated for the tireless efforts put in to make this year’s conference a success

Professor E. R. Adagunodo
Department of Computer Science & Engineering,
Obafemi Awolowo University, Ile-Ife.

LIST OF REVIEWERS

S/No.	Name of Reviewers	Contact Address
1.	Prof. E.R. Adagunodo	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
2.	Prof. G.A. Aderounmu	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
3.	Prof. K. Gbolagade	Department of Computer Science, Kwara State University, Ilorin, Kwara State
4.	Prof. A.O. Oluwatope	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
5.	Prof. B.S. Afolabi	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
6.	Prof. A.I. Oluwaranti	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
7.	Prof. A.O. Odejobi	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
8.	Dr. (Mrs.) R.N. Ikono	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
9.	Dr. K.I. Ogundoyin	Department of Computer Science, Osun State University, Osogbo
10.	Dr. A.O. Ajayi	University of the West of England, United Kingdom (UK) & Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
11.	Dr. A. O. Afolabi	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
12.	Dr. P.A. Idowu	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
13.	Dr. (Mrs.) A.R. Iyanda	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
14.	Dr. S. Aina	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
15.	Dr. Samara Bright	Department of Computer Science, Benue State University, Benue
16.	Dr. K.C. Olufokunbi	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
17.	Dr. (Mrs.) G.O. Binuyo	African Institute for Science Policy and Innovation, Obafemi Awolowo University, Ile-Ife.
18.	Dr. F.O. Asahiah	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
19.	Dr. (Mrs.) S.A. Bello	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
20.	Dr. S. Aina	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife

S/No.	Name of Reviewers	Contact Address
21.	Dr. I.P. Gambo	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
22.	Dr. B.I. Akhigbe	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
23.	Dr. Yange	Department of Computer Science, Federal University of Agriculture, Markurdi, Benue State
24.	Dr. A.O. Adejuwon	African Institute of Science, Policy and Innovation, Obafemi Awolowo University, Ile-Ife
25.	Dr. I.O. Awoyelu	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
26.	Dr. O.D. Ninan	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife
27.	Dr. F.O. Asahiah	Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife

TABLE OF CONTENTS

NOTE

This page was deliberately left blank and to come first before the papers in this proceedings

SECONDARY SCHOOLS STUDENTS' PERCEPTION ON ONLINE INTERACTION AND ICT TRAINING FOR SUSTAINABILITY

[1]Umar, A. K. & [2]Lawal B.
[1]Department of Mathematics and Computer Science, UMYU, Katsina, +2348064174845
[2]Department of Computer Science, FCE, Katsina, +2348038278822

ABSTRACT

Information Communication Technology in Education is one of the key elements around the world that transformed the traditional formal education into modern one. Students do interact with their teachers virtually as they are in the same building, because they share ideas inform of internet forum with their colleagues for educational purpose especially if they are competent in ICT skills. The study was purported to determine the level of perceptions of Science Secondary School Students on online interaction and ICT training. One research question and one hypothesis were formulated. The study used survey research design with sample size of hundred Science Secondary School students selected randomly. A self-structured questionnaire was used with fourteen items. Data was analyzed using simple percentage and SPSS package (One-Sample T-test) and to test the null hypothesis. The result of the null hypothesis was 0.159 at 0.05 level of significance. Therefore, the null hypothesis was accepted. It was found that, in nine out of fourteen items, students were not competent in online interaction while highly competent in only five items not for educational purposes. The paper recommended some solutions to overcome the issue and a way to improve the use of online interaction and ICT skills for instructional purposes.

Keywords: ICT, students' perception, online-interaction, internet training, digital divide, ICT skills

1.0 INTRODUCTION

In recent years, Information Communication Technology (ICT) has taken an important role in the society and are used in a multitude of activities. ICT is already part of most sectors: education, robotics, public administration, employment, business and health. The use of Information Technology and Communication (ICT), leading to the current information society, represents one of the most refreshing proposals for current education systems, through which key project elements around the transformation of various processes in traditional formal education (Galindo, 2011).

The technological advance developed in the last two decades in terms of communication / information, makes clear that social life is changing significantly the degree of also get involved in educational institutions despite the strength of its integrated systems on them (Galindo, 2011). As noted by Bryman (2004: 202) the use of secondary data allows us to make a journey exploring what other researchers have found, comparing and contrasting their findings to see the whole picture of what they have found.

In this article, the paper particularly deals with gaps in competencies or skills using ICT, analyzing the factors that explain the skills and characteristics that may determine different levels of ICT competencies. The main concern of this work is that most students are conversant with online interaction at different capacities, but most of their perceptions to the use of these tools are to the negative, so they hardly give focus on how to affectively use it to develop their capacities and potentialities to the positive. Hence the need for the current work.

The purpose of the write-up is to determine the level of perceptions Secondary School Students have in relation to positive online interaction and ICT training for sustainability in academics and general development. The study is out to answer the question that says:

> *What are the perceptions of Science Secondary School Students on online interaction and internet training for sustainability of ICT?*

The work is delimited to only Science Secondary School Students, only public Schools in Katsina State were captured by the study.

2.0 LITERATURE REVIEW

2.1 ICT and Secondary Education

The case of ICT also raises an issue of particular relevance: the so-called digital divide. This term is used when considering the differences between different groups of people, in their knowledge and mastery of new technologies. These differences may be influenced by socioeconomic factors (for example, there is strong contrast between the developed countries and third world), or other issues such as age and gender. Regarding the latter, it may be of interest exposed by Prensky 2001, who speaks of the natives and digital immigrants. So, one can say that Information and Communications Technology (ICT) is an educational tool unprecedented (Pantoja, & Huertas, 2010).

As mentioned by Bransford, Brown, and Cocking (cited in Huffaker, 2003, p. 357) an active learning allows the students to decide when they require a particular information and whether they have already understood that information or not. This active learning also implies an independent learning. By having access to internet in their school the students will not totally depend on the teachers. They can explore information available in the internet, find information that they need, copy it, and go on to find more and more information. By using this learning system, the students also becomes self-managed in their learning process.

Never before, the Secondary School Students had received volumes of information, however such information is not equal to knowledge, ICT only can help to improve education of students if teachers know how to take such advantage. The problem is till today, the ICT is an underutilized resource in teaching unknown to most teachers that their integration could open the door to a new era of education. This is proven that ICT has only come to the classroom, but it points the way to a profound transformation of the educational model that will involve both students and teachers (Pantoja, & Huertas, 2010). This view has been improved considering the gap in years. In today's classrooms, ICT tools are used by both teachers and students for leisure or various social issues, but not for educational purpose. The current paper is out to examine the perceptions of students to that effect and design a way to improve its use for instructional purposes. Therefore, an increasing number of countries have accepted the need to introduce compulsory education in a formative dimension that provides young people with the necessary keys to understand the technology.

2.2 Factors Explaining the Gaps in ICT Skills

Digital Divide concept is not only related to ICT access, but also with the ability to use these technologies; i.e., skills or abilities that the population needs to acquire for the use of ICT. The effective use of ICT in different areas such as: entertainment, communication, education, etc. (Matamala, 2015). The gap is not only limited to physical access, but also to how people use ICT. Such gaps, is what has been called gaps second order (Matamala, 2015), so this refers to the proper use of ICT in all areas. This is yet another concern of this study. To clear the gap by ensuring effective and efficient accessibility, affordability and usability of ICT tools in the classrooms.

Generational changes have shown that not all human beings are able to incorporate the order of the material discourse that build ICT: they are not just teachers, to name a collective considered central to our society who refuse to arrive at use and application of ICT in the classroom or your life. It has been already shown that generational changes do not always get along with technologies, especially if they affect significantly on the processes of socialization and training of human beings (León & Caudillo, 2014). Today in Katsina, the digital divide is made up of about 70% of the total population with large asymmetries depending of ICT penetration in urban and rural areas; whereas in 630 major urban areas 30% of its population has Internet access in rural areas only 6% of its population (5.9 million households) have a computer and 3% are connected to Internet. It can be established that there are gaps in ICT competencies of secondary school students as socioeconomic level, years of computer use, frequency of computer use and level of confidence in the use of computers, coinciding with the factors that have been identified in previous studies about gaps in the use of ICT by Matamala, 2015.

3.0 METHODOLOGY

The research is an opinion seeking type, so it used a descriptive survey design. Only a sample of one hundred (100) Science Secondary School students (both junior and senior sections) were randomly selected from 2 Schools out of 13 Schools for the study. The research used a self-structured quantitative questionnaire with only fourteen item statements. The instruments was administered by self with the help of a research assistants. Opinions of respondents were analyzed using simple percentages and SPSS package while hypothesis was tested using One-sample T-test to accept or reject null hypothesis.

Table 1: Sample Size of Students by School Sections

Name of School	**Section**		
	Junior	**Senior**	**Total**
Government Science Secondary School, Batagarawa, Katsina	26	24	50
Government Girls Science Secondary School, Ajiwa, Katsina	20	30	50
Total	**46**	**54**	**100**

4.0 RESULTS AND DISCUSSION

4.1 Online Interaction

Table 2 has details of students' perceptions on online interactions and ICT training. In Table 2, the results about online interaction were shown. By type of activities used by students regarding the dimension of online interaction in the first item related to use of e-mail and virtual forum to exchange academic views with peers and teachers, the perception of the students about the competencies level, showed 60.8% consider themselves as not competent, meanwhile 39.2% of the students consider themselves as competent students.

Referring to item 2, use social networks to exchange academic information considered 57.2% students to be not competent and 42.8% students as competent.

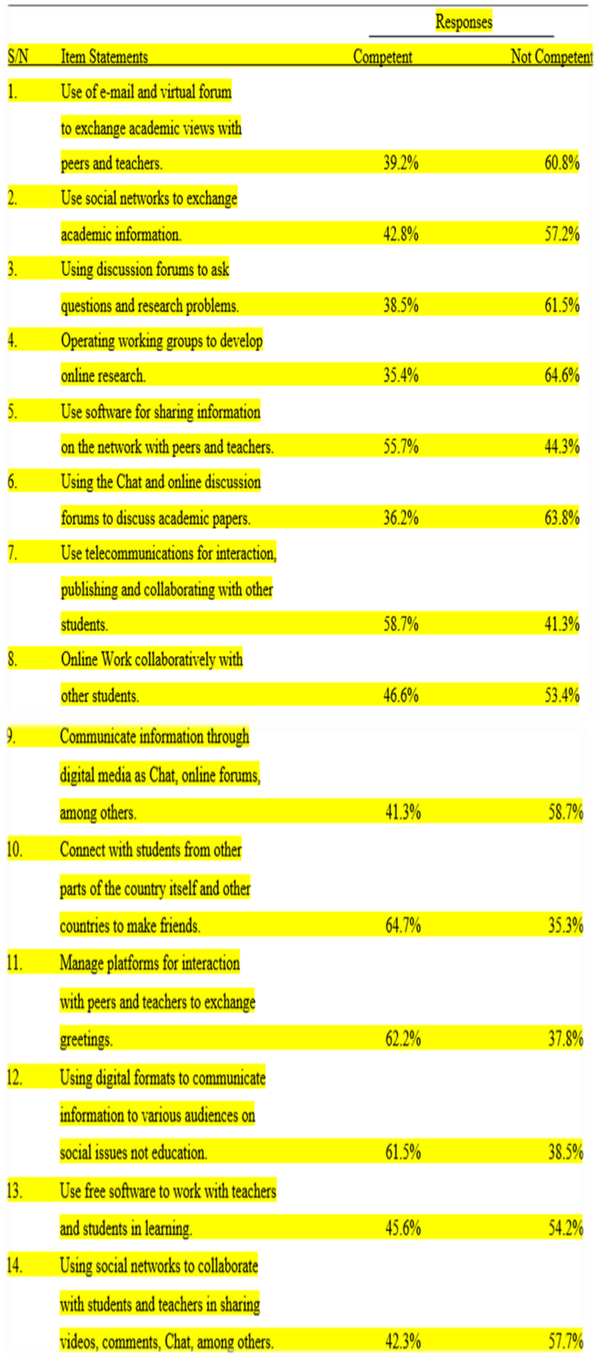
In item 3, using discussion forums to ask questions and research problems, was selected by 38.5% students as competent and 61.5% of students are considered as not competent.

On the other hand, operating working groups to develop online research, in terms of the range of percentages as in the previous case the results are very similar with 35.4% students consider themselves as competent and 64.6% of students consider themselves as not competent students.

Referring to use software for sharing information on the network with peers and teachers, the result obtained was more balanced by 55.7% students consider themselves as competent and 44.3% of students consider themselves as not competent students.

Regarding using the Chat and online discussion forums to discuss academic papers, the results showed 36.2% students as competent 63.8% of students as competent students.

Table 2: Competencies On-Line Interaction



S/N	Item Statements	Responses	
		Competent	Not Competent
1.	Use of e-mail and virtual forum to exchange academic views with peers and teachers.	39.2%	60.8%
2.	Use social networks to exchange academic information.	42.8%	57.2%
3.	Using discussion forums to ask questions and research problems.	38.5%	61.5%
4.	Operating working groups to develop online research.	35.4%	64.6%
5.	Use software for sharing information on the network with peers and teachers.	55.7%	44.3%
6.	Using the Chat and online discussion forums to discuss academic papers.	36.2%	63.8%
7.	Use telecommunications for interaction, publishing and collaborating with other students.	58.7%	41.3%
8.	Online Work collaboratively with other students.	46.6%	53.4%
9.	Communicate information through digital media as Chat, online forums, among others.	41.3%	58.7%
10.	Connect with students from other parts of the country itself and other countries to make friends.	64.7%	35.3%
11.	Manage platforms for interaction with peers and teachers to exchange greetings.	62.2%	37.8%
12.	Using digital formats to communicate information to various audiences on social issues not education.	61.5%	38.5%
13.	Use free software to work with teachers and students in learning.	45.6%	54.2%
14.	Using social networks to collaborate with students and teachers in sharing videos, comments, Chat, among others.	42.3%	57.7%

In item 7, use telecommunications for interaction, publishing and collaborating with other students, an incidence of 58.7% students consider themselves as competent and 41.3% of students consider themselves as not competent students.

As to online Work collaboratively with other students, the frequency of percentages presents results of 46.6% students as competent and 53.4% of students consider themselves as not competent students.

In the case of communicate information through digital media as Chat, online forums, among others, the range of percentages exhibits behavior of 41.3% students consider themselves as competent and 58.7% of students consider themselves as not competent students.

In item 10, connect with students from other parts of the country itself and other countries to make friends, showing a full turn to all results, with 64.7% students consider themselves as competent and 35.3% of students as not competent students.

For item 11, manage platforms for interaction with peers and teachers to exchange greetings, 62.2% students consider themselves as competent and 37.8% of students consider themselves as not competent students.

About item 12, using digital formats to communicate information to various audiences on social issues not education, 61.5% students consider themselves as competent and 38.5% of students consider themselves as not competent students. In item 13, use free software to work with teachers and students in learning, was selected by 45.6% students as competent and 54.2% of students consider themselves as not competent students.

Finally, item 14, that is using social networks to collaborate with students and teachers in sharing videos, comments, Chat, among others, presents results to the frequency on the above results of 42.3% students consider themselves as competent and 57.7% of students consider themselves as not competent students.

However, From the findings of the study, it can be concluded that, Science Secondary School Students in Katsina State are seen to be weak and majority are not competent in such areas as use of e-mail and virtual forum to exchange academic views with peers and teachers, use of social network, use of discussion forums to ask questions and collaborative efforts to develop research online.

Another findings revealed that, incompetency of majority of the students despite being science oriented are in the areas of inability to use software for sharing information on network, using the Chat and online discussion forums to discuss school issues, communicate information through digital media as Chat, online forums with peers, other students and teachers. This is not a good signal, hence something has to be done to arrest these situations.

It can also be seen from the perceptions of the students that, most Secondary School students in Science School in Katsina State were found to be competent only in connecting with other students in other countries to make friends, they are also found to manage platforms for interaction and with peers and teachers to exchange greetings, using digital formats to communicate information to various audiences on social issues not education.

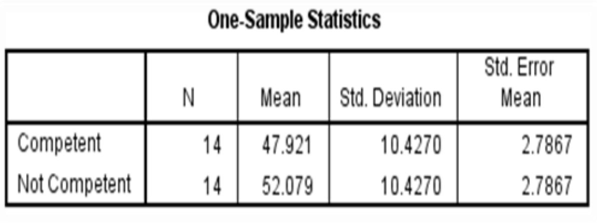
The study is also designed with the following null hypothesis H0 and alternative hypothesis H1 in order to further confirm the reliability of the outcome of this research;

H_0: There is no significant difference in the perceptions of Science Secondary School Students on online interaction and ICT training for sustainability.

H_1: There is a significant difference in the perceptions of Science Secondary School Students on online interaction and ICT training for sustainability

The percentages of the Table 2 (competences on-line interaction) were used to test the hypothesis. The details of the one-sample statistics and t-test carried out is shown in tables 3 and 4 respectively.

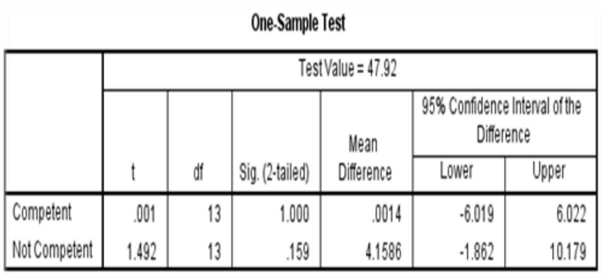
Table 3: One-Sample Statistics



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Competent	14	47.921	10.4270	2.7867
Not Competent	14	52.079	10.4270	2.7867

Table 4: One-Sample T-test carried out



One-Sample Test

	Test Value = 47.92					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Competent	.001	13	1.000	.0014	-6.019	6.022
Not Competent	1.492	13	.159	4.1586	-1.862	10.179

4.2 Interpretation of One-Sample Test

Adopting the one-sample t-test techniques at level of significant = 0.05 and calculated t = 1.64. The t value (t-calculated) was obtained to be 1.49, the degree of freedom (N) = 14, p-value for 2-tailed = 0.159, the differences between true and assumed mean = 4.16. Using the standard of one-sample t-test, we accept the null hypothesis if the calculated t is less than the value of t-critical, and reject the null hypothesis otherwise, also, the null hypothesis is accepted if p-value is greater than the level of significant value and rejected otherwise. Since the t value obtained is less than the t-critical and p-value is greater than level of significant value, then we accept the null hypothesis which states that there is no significant difference in the perceptions of Science Secondary School Students on online interaction and ICT training for sustainability.

5.0 CONCLUSIONS

Today ICT is fundamental to improving the quality of teaching tools, but only if students know how to take advantage, have the proper training and have the necessary resources. Children and adolescents who currently enters educational institutions were born in the digital age; in it, the development of ICT has led to the emergence of novel communication styles and cognitive skills, facilitating the creation of new dimensions in the categories of author and reader that have enabled the construction of a new subject of knowledge (Navés, 2015). There is a low level of ICT competencies in the students of secondary schools in the city of Katsina. This could be because most of the students are not familiar with the tool, the school is short or having lack of such facilities and at some other time, even the teacher have no technical ability to operate the tools.

Apparently the students are in the process of developing these skills. However, they require an increase in frequency and forms of ICT used for academic purposes, greater interaction between peers and teachers in the exchange of views and dissolving doubts, as well as increasing the use of educational platforms, blogs, web pages, etc., for the publication and exchange of educational content.

In this sense, it was found that, students still do not have a high level in the use of ICT for education purposes, which implies a deficiency in the cost-benefit ratio for society. Technological change globally has become a paradigm that appears to regulate the growth of countries; the level of ICT use in education represents a great opportunity for individuals to the path of knowledge and its inclusion in the Current society characterized by a self-learning management supported by the application of digital skill.

REFERENCES

Pantoja, A. & Huertas, A. (2010). Integracion de las TIC en la asignatura de tecnologia de educacion secundaria. *Pixel-Bit Revista de Medios y Educación, 37,* 225-337. Recovered from: http://www.sav.us.es/pixelbit/pixelbit/articulos/n37/18.pdf

Matamala, C. (2015). Factores predictivos de las competencias TIC en alumnos chilenos de secundaria. *Revista Iberoamericana de Educación, 67 (1),* 121-136. Recovered from: http://www.rieoei.org/deloslectores/6701Matamala.pdf

León, G. A. & Caudillo, D. (2014). Relaciones interactivas, internet y jóvenes de secundaria en México: primera oleada sobre usos, consumos, competencias y navegacion segura en internet en Sonora(2013). *Revista internacional de tecnologías en la educación, 1 (2),* 41-50. Recovered from: http://sobrelaeducacion.com/_uploads/Rev_Int_de_Tecnologias_en_la_Educacion_1%282%29,_2014.pdf

Crovi, D. M. (2010). Jóvenes, migraciones digitales y brecha tecnologica. *Revista mexicana de ciencias politicas y sociales, 52 (209),* 119-133. Recovered from:http://www.revistas.unam.mx/index.php/rmspys/article/view/25967

Navés, F. A. (Enero-Junio 2015). Las TIC como recurso didáctico:¿Competencias o posición subjetiva? *Revista de Investigacion Educativa, 20,* 238-248. Recovered from: http://revistas.uv.mx/index.php/cpue/article/view/1308/2399

Galindo, A. C. (2011). Representaciones sociales de las prácticas tecno-educativas en profesores de nivel basico (secundaria) en distintos contextos educativos en Colima. Congreso Internacional Educación Mediática y Competencia Digital 2011. Recovered from: http://www.educacionmediatica.es/comunicaciones/Eje%201/Alma%20Celia%20Galindo.pdf

Adi Suryani (2010). Ict In Education: Its Benefits, Difficulties, And Organizational Development Issues, jsh Jurnal Sosial Humaniora, Vol 3 No.1, Juni 2010.

Huffaker, D. (2003). Reconnecting the classroom: E-learning pedagogy in US public high schools. Australian Journal of Educational Technology, 19 (3), 356-370.

STACKED ENSEMBLE MODEL FOR DIAGNOSIS OF HEAD AND NECK CANCER IN PRIMARY HEALTHCARE SYSTEM

[1]Olabode O. O., [2]Adetunmbi A. O., [3]Akinbohun F. and [4]Ambrose A.
[3]Department of Computer Science, Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria
[1,2]Federal University of Technology, Akure, Ondo State, Nigeria.
[4]Department of Ear, Nose & Throat/Head & Neck, University of Medical Sciences Teaching Hospital,
Akure, Ondo State, Nigeria

*Email of Corresponding Author: *folakeakinbohun@yahoo.com*

ABSTRACT

Head and neck cancers (HNC) are indicated when body's cells begin to divide without stopping and grow out of control around the Head and Neck regions. The increase of the epidemiology of head and neck cancer needs computational intelligent tool for prediction at the peripheral level for prompt referral. Hence, data analytics driven models can be applied in medical domain with the aid of machine learning to predict diseases. The paper considered two feature selection methods namely consistency and chi square methods on the whole dataset obtained from hospitals in South-West, Nigeria. The Stacking Ensemble Method was selected which combined multiple classifications namely Decision tree (C4.5), KNN and Naive Bayes (the base-level classifiers) via a meta-classifier namely logistic regression where cross validation was applied. On the level of ensemble learning, when logistic regression was used at meta-level on the reduced dataset, the result indicated that chi square method in a stacked ensemble model produced better prediction than consistency method. Chi Square feature method on stacked ensemble model can be used for prediction of head and neck cancer. The work was limited to only four types of head and neck cancer.

Keywords: Consistency, decision tree, head and neck cancer, KNN.

1.0 INTRODUCTION

Head and neck cancers are cancers associated with the head and neck regions. They are disfiguring diseases associated with a high mortality rate (Hagedoorn and Molleman, 2006; GBD, 2016). HNCs are the sixth most common cancers worldwide (John *et al*, 2000) with increasing incidence in developing countries.

Head and neck cancer pose a challenge to both the patient and the non-specialists in Otorhinlaryngology (a sub-specialty in medicine that deals with ear, nose, throat/head and neck regions) because of their anatomical sites. Onotai and Nwogbo (2012) carried out a study to draw attention to the recent trends of primary head and neck tumors in Port Harcourt, Nigeria.

The healthcare analytics focuses on retrospective data. It provides useful information in order to make a decision in the area of healthcare. The healthcare system continues its transition to new care delivery model in healthcare organizations where ensemble model is embraced to diagnose cancers in primary healthcare centres. An ensemble model can be created which helps in identifying patients who are likely to have a particular type of cancer around head and neck region (craniocervical). This helps individual and health workers in both primary and secondary healthcare centres where there are no head and neck experts to diagnose Head and Neck Cancer (HNC). Ensemble systems have proven effective in computational intelligence and machine learning. Ensemble methods are developed to reduce bias, variance and thereby improve the accuracy of an automated decision-making system. Ensemble systems have since been successfully used to address a variety of machine learning tasks of which to diagnose the types of head and neck cancer is not exempted

In medical domain, where data analytics-driven models are successfully applied, new directions are identified to predict diseases. In order to make the models predict diseases, machine learning methods are involved (Hussein *et al*, 2002). Medical diagnosis needs to be executed accurately and efficiently in order to make individuals and government aware of the importance of ensemble systems especially in the area of Head and Neck Cancers. There is a shortage of medical specialists in head and neck sub-specialties in medicine, therefore predictive diagnostic system would be exceedingly beneficial to enable non-specialist doctors and attending paramedics make appropriate and timely diagnosis and prompt referral for further management.

The rate of death from head and neck cancers is high (WHO, 2014) yet it is overwhelming that late presentation to the specialists is common especially in the third world countries. This paper uses learning algorithms to diagnose patients with HNC having considered their clinical features.

Patients' delay in consulting a health care professional is attributed to patient delay in recognizing the signs and symptoms of cancer or referring to the professional or specialist in head and neck (ENT-Ear, Nose and Throat) (Hollows *et al*, 2000) for prompt attention. The delay by some health workers can be attributed to failure in recognizing the clinical features suggestive of head and neck cancers. The use of machine learning algorithms is able to diagnose the existence of head and neck cancers. The objective of this paper is to develop a stacked ensemble model for diagnosing patients with head and neck cancer (HNC).

2.0 RELATED WORK

This section considers the related work of some researchers on head and neck cancer using machine learning algorithms.

Andre et al. (2019) worked on head and neck cancer outcome prediction using deep learning. The authors tested their hypothesis by training a Convolutional Neural Network (CNN) to predict treatment outcomes of patients with head and neck cancer, based solely on their pre-treatment computed tomography image. The results got included AUC of 0.88 in predicting distant metastasis. The work only used CNN

Martin *et al*. (2018) developed a tissue classifier using three distinct convolutional neural network (CNN) architectures on Hyperspectral Imaging (HIS) data to investigate the ability to classify the cancer margins from *ex-vivo* human surgical specimens. The CNN-based method classifies the tumor-normal margin of cancer versus normal oral tissue with an area under the curve (AUC) of 0.86 for inter-patient validation, performing with 81% accuracy, 84% sensitivity, and 77% specificity. Thyroid carcinoma cancer-normal margins are classified with an AUC of 0.94 for inter-patient validation, performing with 90% accuracy, 91% sensitivity, and 88% specificity.

Amanda *et al.* (2018) researched on evaluation of classification and regression tree (CART) model in weight loss prediction following head and neck cancer radiation therapy. The result showed that the area under the curve during radiation therapy (RT) and End of Treatment (EOT) were 0.773 and 0.821 respectively. The potential value of an informatics has facilitated insight into the prediction of weight loss using the CART algorithm was established. The limitations of the study was that only one model (CART) was used for prediction.

A model to predict critical weight loss in patients with head and neck cancer during chemotherapy and radiotherapy (CRT) was presented by Jacqueline *et al.* (2016). Logistic regression and classification and regression tree (CART) analyses were used to analyse predictive factors for critical weight loss (defined as >5%) during (C)RT. Fifty percent of patients had critical weight loss during (C)RT. The main predictors for critical weight loss during (C)RT by both logistic and CART analyses were RT on the lymph nodes, higher RT dose on the primary tumour. The model predicted radiotherapy critical weight loss in patients with head and neck cancer. Two models (Logistic regression and classification and regression tree (CART) were used for the study.

Basma *et al*. (2016) titled a paper "Performance of Data Mining Techniques to Predict in Healthcare Case Study in Chronic Kidney Failure Disease". The objective of this research was to predict kidney disease by using multiple machine learning algorithms. Three learning algorithms on a set of medical data were applied. They were Support Vector Machine (SVM), Decision Tree (C4.5), and Bayesian Network (BN), and chose the most efficient one. Contribution to knowledge was that the work was able to use multiple learning machine algorithms to predict chronic kidney failure disease. The study was limited to the inability to deploy feature selection method on the dataset and only one disease (Chronic kidney disease) was considered.

Sami *et al.* (2015) proposed prediction on depression and quality of life (QOL) among long-term head and neck cancer survivors. Significant predictors of fatigue, global health/QOL, social contact, speech, pain, swallowing, and xerostomia were identified. A total of 209 post treatment (median follow-up, 38.7 months) head and neck cancer patients were prospectively evaluated. T-test and linear regression analyses were used to construct predictive models. Pretreatment predictors of long-term depression and QOL have been defined using multivariate models. Small sample size was used and a few numbers of predictors/features were considered

A study on predictive data mining for diagnosis of thyroid disease using neural network was developed by Prerana *et al* (2015). The paper presented a systematic approach for earlier diagnosis of thyroid disease using back propagation algorithm. The result of the study showed that outcome of Artificial Neural Network (ANN) was in good agreement with experimental data. The limitations of the study were that only one algorithm was used (Neural network). There was no room to compare other algorithms. The work diagnoses only one kind of disease (Thyroid).

Durairaj and Deepika (2015) titled a paper prediction of acute myeloid leukemia cancer using data mining-a survey. Data mining techniques such as Bayes Network, JRip, J48, Multilayer perceptron, IBK, Decision Tree were used on the dataset.

Dietterich (2000) proposed ensemble methods in machine learning. The paper reviewed ensemble algorithms such as bayesian averaging, error-correcting output coding, bagging, and boosting. This paper reviewed these methods and explained why ensembles could often perform better than any single classifier.

The annual incidence of head and neck cancers worldwide is more than 550,000 cases with around 300,000 deaths each year (Jemal *et al*., 2011).

Baatenburg *et al.* (2001) proposed prediction of survival in patients with head and neck cancer. Prognostic value of site of the primary tumor, age at diagnosis, gender, T-, N-, and M-stage, and prior malignancies of 1396 patients were studied univariately by Kaplan-Meier curves and the log-rank test. The Cox-regression model was used to investigate the effect of these variables simultaneously on a prediction model of survival in individual patients.

Allgar and Neal (2005) worked on diagnostic delays that occurred because of late patient presentation and problems at the primary health care. The delay result in poor prognosis.

Opitz and Maclin (1999) looked into the empirical study of two ensemble methods. They used bagging and Boosting methods on 23 data sets using both neural networks and decision trees as their classification algorithms. It was found out that bagging was almost always more accurate than a single classifier, and sometimes, much less accurate than Boosting. Boosting ensemble results were less accurate than a single classifier especially when using neural networks. Further results showed that Boosting ensembles might overfit noisy data sets, thus decreasing its performance. Limitation of the research was that stacking ensemble method was not part of the study.

3.0 METHODOLOGY

The development of stacked ensemble models on diagnosis of head and neck cancers involves some stages such as data preprocessing (data selection, data cleaning, replacing of missing values and discretizing attributes), feature selection, learning algorithms and stacked ensemble models. The architecture of a stacked ensemble model for head and neck cancer diagnosis is resented in figure 1.

A Data Collection

The raw data used in this research were obtained from University of Medical Sciences Teaching Hospital, Akure and Federal Medical Centre, Owo. The collected data in Pathology Departments contain all forms of cancers in human body, in which Head and Neck Cancers were selected by Head and Neck surgeons. The selected data were traced out to patients' case notes to identify the clinical features.

B Preprocessing

In this stage, data cleaning and discretization were carried out.

The Structure of the dataset

Table 1 contains features or attributes and their types. Data types are represented with the following data values. 0 and 1 represent no and yes. While 2, 3, 4 to represent mild, moderate and severe; small/medium/ large represent 1, 2 and 3. HNC data set consists of 1473 instances and 18 features.

Table 1: HNC Data Set

S/No.	Feature/predictors variable name	Predictors type
1	Bleeding	Mild/moderate/severe/ (absent)
2	Poor appetite (Anorexia)	Yes/no
3	Weight loss	Yes/no
4	Snoring	Yes/no
5	Swelling	Small/medium/large
6	Nasal blockage	Yes/no
7	Mouth breathing	Yes/no
8	Hyponasal speech	Yes/no
9	Halithosis	Yes/no
10	Facial Asymmetry	Yes/no
11	Fatigue	Yes/no
12	Hoarseness	Yes/no
13	Dyspnoea	Yes/no
14	Tinnitus	Yes/no
15	Haemoptosis	Yes/no
16	Proptosis	Yes/no
17	Odynophagia	Yes/no
18	Dysphagia	Yes/no
	Class- Cancer type	Sinonasal/nasopharyngeal/ laryngeal/ thyroid

The description of the classes are as follows:

Thyroid cancer: The cancer that affects the front part (anterior) of the neck.

Nasopharyngeal cancer: This is the most common malignant tumor (cancer) of the nasopharynx. It is a cancer that starts in the nasopharynx, the upper part of the throat behind the nose and near the base of skull.

Laryngeal cancer: It affects the larynx which is commonly called the voice box. It is located at the top of the windpipe or trachea

Sinonasal cancer: Sinonasal tumors are tumors that occur in the nasal cavity or nasal sinuses. The nasal cavity is otherwise called the nostril where air passes on its way to the lung.

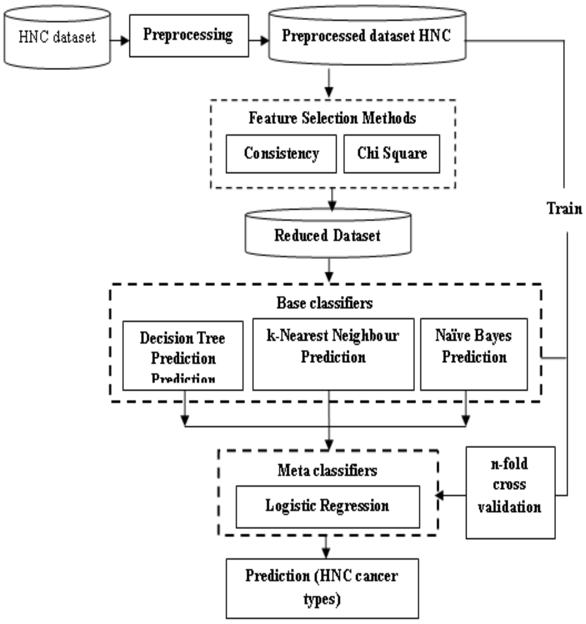


Figure .1: Architecture of the Proposed Stacked Ensemble for HNC diagnosis model

C. Feature Selection

There is need for feature selection method to remove redundant and irrelevant features (Jiliang *et al* 2000). It is important to consider feature selection a part of the model selection process in order to remove bias from the models; and improve accuracy and reduce training time. Consistency method and chi square method were adopted during this stage.

D. Model Construction (Stacked)

Another step to follow after feature selection is the selection of modeling techniques. Stacking was selected which combined multiple classifications (the base-level classifiers) via a meta-classifier. To use stacking ensemble method, various learning algorithms L_1, L_2, L_3,…L_n are combined on a single dataset, D which consists of $D = (x_i, y_i)$. In stacking ensemble learning, a phase uses a set of base-level classifiers C_1, C_2, …, C_n which are generated where each classifier is learnt on the dataset i.e.

$C_n = L_n (D)$.

First phase: Base Level Classifier

Three algorithms namely Decision Tree (C4.5), Naïve Bayes and kNN are selected as base level classifier.

(I) Decision Tree (DT) (C4.5)

Decision tree uses greedy algorithm which is constructed in a top-down recursive divide-and-conquer manner. A decision tree is used as a classifier for determining an appropriate action for a given case (Jiawei *et al.*, 2011). It diagnoses or determines the type of cancer in the head and neck region a patient could have. Information about a patient is given as vectors of attributes or input variables. To find the appropriate class for a given patient (a person), it starts with the test at the root of the tree and keep following the branches as determined by the values of the features of the case at hand, until a leaf is reached.

Decision Tree (C4.5) uses information gain where entropy for each branch is calculated (i.e. the entropy of the class and each subset of the attribute/feature) are given in Equation 1. The algorithm (C4.5) uses Gain Ratio which is computed using Equation 2

$$\mathrm{E} = -\sum_{i=1}^{n} \mathrm{p_i}\, \boldsymbol{log_2 p_i} \qquad (1)$$

Where Pi is the proportion of examples in hepatitis that belong to the i-th class, n is number of classes.

$$\text{Gain Ratio}_{\text{attribute}} = \frac{\text{Gain}_{\text{attribute}}}{\text{SplitInfo}_{\text{attribute}}} \qquad (2)$$

(II) Naïve Bayes Classifier

Naïve Bayes algorithm is a classification technique that works on Bayes theorem of probability to predict the class of unknown dataset. Using conditional probability model of Naïve Bayes, a problem instance to be classified is represented by a vector x=(x_1, x_2, x_3, … x_n) representing features (independent variables) and it assigns to this instance probabilities as follows: Pr(C_m| x_1, x_2, x_3, … x_n). The Bayes Theorem is as stated in Equation 3.

$$\mathrm{P}(c_m|\boldsymbol{x}) = \frac{\mathrm{P}(\mathrm{x}|\mathrm{C_m}) * \boldsymbol{P}(\mathrm{C_m})}{\mathrm{P}(\mathrm{x})} \qquad (3)$$

To construct a classifier from the conditional probability model, Bayes classifier combines the model with a decision rule called Maximum A Posteriori (MAP). ((Jiawei *et al*, 2011)

P(c_m|x) is the posterior probability of class (c, HNC type) given predictor (x, features)

P(c_m) is the prior probability of class

P(x|c_m) is the likelihood which is the probability of predictor given class

P(x) is the prior probability of predictor

The prior probability is the evidence, and evidence is always 1. It is used to normalize the result. It remains the same, removing it will not have effect on the prediction.

(III) K-Nearest Neighbor (KNN)

K-Nearest Neighbor is an algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g. distance functions). When a prediction is required for unknown data instance, the KNN algorithm will search through the training dataset for the k-most similar instances. The Euclidean distance between a training tuple $[\!(\mathrm{ID}]\!]\_\mathrm{xi}$ and a query (given) tuple $\mathrm{ID_x}$ to be classified can be derived in Equation 4:

$$\mathrm{Euclidean} = \sqrt{\sum_{i}^{n} [\!(\mathrm{ID}]\!]_{xi} - \mathrm{ID_x})^2} \qquad (4)$$

Second phase: Meta level classifier

The predictions from the base classifiers are combined. These predictions were passed to the meta dataset and a meta algorithm was used. The next phase makes use of regression algorithm at the meta-level classifier to be learnt which combines the outputs of the base-level classifiers. The logistic regression uses a linear predictor function *f*(c,y) to predict the probability that observation *y* has outcome c, which is given by Equation 5

$$\mathrm{f(c,y)} = \boldsymbol{\beta}_{0,c} + \beta_{1,c}x_{1,y} + \beta_{2,c}x_{2,y} + \dots + \beta_{m,c}x_{m,y} \qquad (5)$$

where $\beta_{m,c}$ $\beta_{m,c}$ represents regression coefficient associated with the mth explanatory variable and the cth outcome.

At the meta-level classification, the meta-level dataset consists of the form: $\bar{Y}^1_{n,} \dots \bar{Y}^m_{n,}$ a training set is generated where cross validation step is applied. In the procedure for cross validation, each of the base-level algorithms is applied to almost the entire dataset leaving one part for testing: C^i_n = L_n (D-d_i). The learned classifier is used to generate prediction.

4.0 RESULTS AND DISCUSSION

Following the description and preprocessing of the HNC dataset, feature selection methods used were consistency based and chi square method. Table 2 shows the results of the consistency method and chi square

Table 2: Features selected by consistency based and chi square

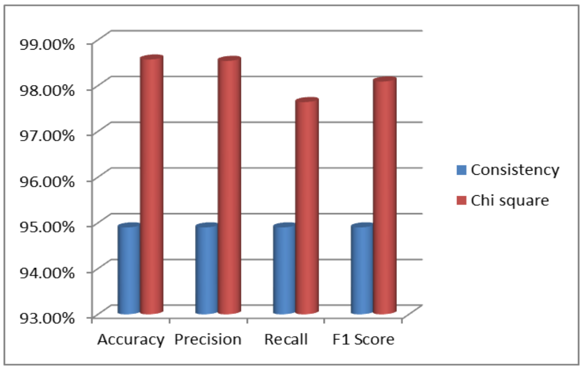
Consistency	Chi square
bleeding	dyspnoea
facial asymmetry	haemoptysis
fatigue	nasal blockage
snoring	hoarseness
hoarseness	snoring
haemoptysis	fatigue
proptosis	poor appetite
dyspnoea	odynophagia
nasal blockage	facial asymmetry
	tinnitus
	hyponasal speech
	mouth breathing
	proptosis
	halithosis
	dysphagia

Chi Square method selected 15 attributes while 9 features were selected by consistency. The number of features selected by chi square were higher than the number of features selected by Consistency method. The features with higher values were selected in the course of using chi square method while features with lower chi square values were cut off (weight loss and bleeding and swelling). The selected features by chi square had good relationship between the features and the class. The 9 features selected by Consistency method indicated that the features were consistent with the class/target.

The selected features were trained using three classification algorithms: Decision Tree (C45), Naïve Bayes and KNN. These were base algorithms/classifiers. the predictions of the base learners that were combined, Logistic Regression algorithm was used at stacked meta level to train the dataset in a stratified cross validation. This was done on the two feature selection methods using consistency and chi square methods. This stacked model was performed to achieve the best classification accuracy; the prediction performance evaluation is presented in Table 3 based on the selected features of Consistency and Chi square methods.

Table 3: Performance evaluation of a stacked ensemble with logistic regression on consistency and chi square methods

Metrics	**Consistency**	**Chi square**
Accuracy	94.90%	98.57%
Precision	0.949	0.9854
Recall	0.949	0.9764
F1 Score	0.949	0.9809
Kappa statistic	0.9309	0.9803
Mean absolute error	0.0485	0.0161



Using Logistic Regression on the selected features by Consistency method, the result showed that the accuracy was 94.90% while chi square produced classification accuracy of 98.57%. This means that Chi Square feature selection method had higher accuracy than Consistency feature selection method. F1 score (being a good metric that evaluates the harmonic mean of precision and recall) had 0.9809 when chi square method was adopted; hence this showed a good measure which determines the best model. When using Consistency method, F1 Score was 0.949 which was not up to the F1 Score of Chi Square. This indicates that using Logistic Regression with chi square method had better performance evaluation than Consistency method. The kappa statistic for both Consistency and Chi Square methods were 0.9309 and 0.9803 respectively, meaning that there were perfect agreement between the observed and the expected values between the two considered feature selection methods. From Figure 2, when considering the Consistency method, performance metrics such as precision, recall F1 Score and accuracy) were not as that of Chi Square. This may be as a result of how Consistency method performed its selection which was done in a batch selective method. Another point to note is that, when using Consistency method, the accuracy and F1 Score were greater than 90%. This showed a good performance. Considering Logistic Regression with Chi Square feature selection method was better than Consistency method. This yielded a better stacked ensemble result that produces the type of cancer in head and neck region.

CONCLUSION

A way of analyzing data through prediction is by using machine learning tools. Two feature selection methods were used to remove noise from the data. Three base models such as Decision tree, KNN and Naive Bayes were used and logistic regression was considered for the meta algorithm. When using logistic regression at meta level, it produced 94.90% accuracy on consistency feature selection and 98.57% classification accuracy on chi square feature selection. It is obvious that chi square method had a higher accuracy. This may be as a result of choice of application of feature selection method which removed irrelevant features. At the end of this study, a stacked ensemble model was developed which can be used in predicting the cancer type around head and neck region. Hence, algorithmic tools can be used to predict the type of head and neck cancer which can assist in complementing clinical decisions. Future work can include other class in head and neck cancer (cancer type around head and neck regions) such as oral cavity and oropharyngeal cancers, ear and hypopharyngeal. In addition, other machine learning algorithms can be used on the dataset at both base level and meta level such as neural network. Data used can be obtained from other tertiary health centres in other regions of the country.

REFERENCES

Allgar V. L. and Neal R. D.(2005). Delays in the diagnosis of six cancers: analysis of data from the National Survey of National Health Scheme (NHS) Patients: Cancer. Br J Cancer.

Andre Diamant, Avishek Chatterjee, Martin Vallieres, George Shenouda and Jan Seuntjens (2019). Deep learning in head & neck cancer outcome prediction. Scientific Report. 2764

Baatenburg de Jong R. J., Hermans J., Molenar J., Briaire J. J. and le Cessie S. (2001). Prediction of survival in patients with head and neck cancer. National Center for Biotechnology Information, U. S. National Library of Medicine. John Wiley & Sons, Inc.

Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq (2016). Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease. *International journal of database management systems* (IJDMS) Vol.8, No.3.

Dietterich T.G. (2000) . Ensemble Methods in Machine Learning. Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelb

Durairaj M. and Deepika R (2015). Prediction of acute myeloid leukemia cancer using datamining-a survey. *International Journal of Emerging Technology and Innovative Engineering* Volume I, Issue 2, February 2015 ISSN: 2394 - 6598

GBD (2016). Disease and Injury Incidence and Prevalence, Collaborators. "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015". *Lancet. 388 (10053): 1545–1602. Available from http:www.ncbi.nlm.gov/PMC* /articles/PMC5055577

Hagedoorn M. and Molleman E.(2006). Facial disfigurement in patients with head and neck cancer: the role of social self-efficacy. *Health Psychology*

Jacqueline A. E. Langius, Jos Twisk, Martine Kampman, Patricia Doomaert, Mark H.H. Kramer, PeterJ. M.Weijs, C Rene Leemans (2015). Prediction model to predict critical weight loss in patients with head and neck cancer during (chemo) radiotherapy. Oral Oncology, Volume 52, Pages 91-96

Jiawei Han, Micheline Kamber and Jian Pei (2011). Data mining: concepts and techniques (3rd edition).

Jiliang Tang, Salem Alelyani and Huan Liu (2014). Feature Selection for Classification: A Review

Jemal A., Bray F., Center M.M., Ferlay J., Ward E, Forman . (2011). Global cancer statistics. *CA Cancer J Clin 2011; 61(2):69-90.*

John C W., Mark N. G. and Janet A. W. (2000). Stell and Maran’s Head and Neck Surgery. Reed educational and professional publishing Ltd. Fourth edition.

Martin Halicek, James V. Little, Xu Wang, Mihir Patel, Christopher C. Griffith and Army Y. Chen and Baowei Fei (2018). Tumor Margin Classification of Head and Neck Cancer Using Hyperspectral Imaging and Convolutional Neural Networks. *Proc SPIE Int Soc Opt Eng.*

Opitz David and Maclin Richard (1999). Popular Ensemble Methods: An Empirical Study *Journal of Artificial Intelligence Research* page 169-198.

Onotai L.O. and Nwogbo A.C (2012). Primary head and neck malignant tumours in Port Harcourt, Nigeria: A revisit. *Journal of medicine and medical sciences* vol. 3(2) pp. 122-125

Prerana, Parveen Sehgal, Khushboo Taneja (2015). Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network. *International Journal of Research in Management, Science & Technology* (E-ISSN: 2321-3264) Vol. 3, No. 2. Available at www.ijrmst.org

Sami P. Moubayed, John S. Sampalis, Tareck Ayad, Louis Guertin, Eric Bissada,Olguta E. Gologan, Denis Soulie`res, Louise Lambert, Edith Filion, Phuc Felix Nguyen-Tan, and Apostolos Christopoulos (2015). Predicting depression and quality of life among long-term head and neck cancer survivors. *American Academy of Otolaryngology*—Head and Neck Surgery. Vol. 152(1) 91–97

World Health Organization(2014). World Cancer Report 2014. Chapter 5.8. *ISBN 9283204298.*

DEVELOPMENT OF AN AUTOMATIC SOFTWARE COMPLEXITY MEASUREMENT SYSTEM

***[1]Sanusi B. A., [2]Olabiyisi S. O., [3]Afolabi A. O. and [4]Olowoye A. O.**
Department of Computer Science & Engineering, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

*Email of Corresponding Author: *sanusibashiradewale90@gmail.com*

ABSTRACT
Code Complexity measures can be used to predict critical information about testability, reliability, and maintainability of software systems from the automatic measurement of the source code. In this paper, existing code-based complexity metrics: Source Line of Code, McCabe Cyclomatic Complexity metrics and Halstead Complexity Metrics were studied and implemented in order to extend the existing schemes. The system was built following the procedure of the waterfall model of software engineering. The system was developed in Visual Studio Integrated Development Environment using C-Sharp programming language, .NET framework and MYSQL server for database design. The performance of this system was tested efficiently using a software testing method known as Black-box testing to examine the functionality and quality of the system. The results of the evaluation showed that the system produced functionality of 100%, 100%, 75%, 75%, and 100%, and quality of 100%, 100%, 75%, 75%, and 100% for the source code written in C++, C, Python, C# and JavaScript programming languages respectively. Hence, this tool helps software developers to view the quality of their code in terms of code metrics. Also, all data concerning the measured source code was well documented and stored for maintenance and functionality in the possibility of future development.

Keywords: Software Metrics, Code-based Complexity, Source Line of Code, McCabe Cyclomatic, Halstead Complexity.

1.0 INTRODUCTION

Computational Complexity theory focuses on classifying computational problems according to their inherent difficulty and relating these classes to each other (Sanjeev and Barak, 2009). A computational problem is a task solved by a computer using mechanical application of mathematical steps, such as an algorithm (Sanjeev and Barak, 2009). One of the roles of computational complexity theory is to determine the practical limits on what computers can and cannot do.

Software complexity is one of the most important concerns in software lifetime development. Software complexity metrics are used to quantify a variety of software properties. Code complexity measures can be used to predict critical information about testability, reliability, and maintainability of software systems from the automatic measurement of the source code (Olabiyisi *et al.* 2013). If the code is complex, then the developers will likely face more problems while developing and maintaining it. Recently, there as been the need for software engineering to accurately predict the complexity of source code to save millions of people in maintaining time and effort (Silva *et al.* 2012).

A software metric is a standard of measure of a degree to which a software system or process possesses some property. Even if a metric is not a measurement (metrics are functions, while measurements are the numbers obtained by the application of metrics), often the two terms are used as synonyms. Since quantitative measurements are essential in all sciences, there is a continuous effort by computer science practitioners and theoreticians to bring similar approaches to software development (Lincke *et al.* 2008).

The field of software metrics has grown up out of two needs (Fenton and Bieman, 2014). One is the need for software developers to be able to manage the software development process. For example, developers need to determine the resources or time needed to deliver a product or whether the software is good enough to release. The other need comes from researchers, who want to be able to objectively define and measure software attributes to get a better understanding of software engineering (Fenton and Bieman, 2014). This field of software metrics is constantly changing. Therefore, there is no standard set of metrics, and new measures are always being proposed. Metrics researchers have to modify their existing parser tools to accommodate the new measures. This is a real challenge to developers since such tools usually have a very complex source code (Scotto *et al.* 2004).

Some software development practitioners point out that simplistic measurement can cause more harm than good (Kaner, 2004). Others have noted that metrics have become an integral part of the software development process (Binstock, 2010). Impact of measurement on programmer's psychology have raised concerns for harmful effects on performance due to stress, performance anxiety, and attempts to cheat the metrics, while others find it to have a positive impact on developers' value towards their work, and prevent them from being undervalued. Some argue that the definition of many measurement methodologies is imprecise and consequently, it is often unclear how tools for computing them arrive at a particular result (Lincke *et al.* 2008), while others argue that imperfect quantification is better than none (You cannot control, what you cannot measure.). Evidence shows that software metrics are being widely used by government agencies, NASA, IT consultants, academic institutions.

Hence, this paper aimed at developing an automated system that evaluate the software complexity of C-family programming language as well as compute the defect rate of the code fragment.

2.0 RELATED WORKS

Zhang and Baddoo (2007) studied the performance of three complexity metrics. They selected McCabes Cyclomatic Complexity, Halsteads Complexity, and Douces Spatial Complexity. Their experiment is based on four hypotheses using data from Eclipse JDT which is an open-source application. As a result, they conclude that the three complexity metrics show different performance results during their hypotheses testing, and finally they recommended combining Cyclomatic and Halstead metrics for better judgments on software complexity.

Another study by Panovski (2008) presented a new assessment of software product quality, which focused on assessing the quality of the external features of the software product, which means evaluating the behavior of the software product when implemented. Also, the study focused on the development of the quality model (ISO / IEC 9126) at the level of software metrics. The study relied on seven samples of the software product and evaluated them using ISO / IEC 9126-2 quality model. In the research, Panovski (2008) concluded that external product quality attributes are an area or category that can be adopted and that the metrics provided by ISO / IEC 9126-2 can be considered as a starting point for the definition of standards, but are not ready to use in their present form. The metrics of the software product need to be more adapted to show better information (Panovski, 2008).

Borchert (2008) discussed the method of code profiling by using a static analysis. The study was done on 19 industrial samples and 37 samples of students' programs. He has analyzed software samples through software metrics. The results of this study indicated that the code pattern could be a useful technique for rapid program comparisons and quality observation in the field of industrial application and education.

Jay *et al*. (2009) made a comparison between two metrics to calculate the complexity of code. They used Cyclomatic Complexity (CC) and Line of code (LOC) to prove the stable linear relationship between these two techniques. They used five NASA projects with different programming languages, such as C, C++, and Java, to be the dataset for their research.

Moreover, Bhatti (2010) explored the area occupied by the software metrics. He used a QA-C tool to measure software metrics automatically on the code written in C programming language through expressing the association between software metrics and the complexity of the source code. He attempted to demonstrate the values of these metrics graphically only, without considering the quality features and threshold limits relationship (Bhatti, 2010).

Another work in 2010 is the impact of code complexity and usability, either in monitoring software complexity during development, or in evaluating the complexity of legacy software. The researchers of this work, Widheden and Goran (2010) proposed a new coupling metrics (Ecoup), and introduced the Java met tool, which works in a static analysis of programs written in Java with respect to coupling, flow control, complexity and coherence. In the same year, Chandra *et al*. proposed the use of Object-Oriented metrics that introduced by Chidamber and Kemerer (1994) to assess program quality at the class level. The proposed tool can be used to verify the class design conforms to the design specifications of the Object Oriented programming, through using the threshold for each metric (Chandra *et al*. 2010).

Silva *et al*. (2012) study the applicability of three software complexity metrics: McCabe’s Cyclomatic complexity, Halstead’s complexity and Shao and Wang’s cognitive functional size. In their study, they used ten different programs of the same programming language (Java) and determined which one of the three complexity metrics is most appropriate for software manufacturing. Also, to manually calculate the ranking of the ten programs complexity based on the three metrics, they applied Quota sampling method by selecting five big companies and randomly asked six programmers from each company to rank the complexity of the ten programs.

Silva *et al*. (2013) made a comparison to test relationship between three cognitive complexity metrics: Kushwaha and Misra (KM’s) cognitive information complexity measure (CICM), Shao and Wang’s (SW’s) cognitive functional size (CFS), and Misra’s cognitive weight complexity measure (CWCM).

Jian *et al.* (2017) worked on using Convolutional Neural Network to predict defects in software. This work focused only on file level prediction. Jian *et al.* (2017) claim that programs have well-defined syntax and rich semantics hidden in the Abstract Syntax Trees, which traditional features often fail to capture and the traditional methods does not really give satisfactory results.

Yu *et al.* (2018) then developed a model with a combination of derived metric sets to improve prediction of defect in concurrent software programs using deep learning technique which was named ConPredictor.

Hence, this paper study and investigate code-based complexity metrics, Source Line of Code (SLOC), McCabe Cyclomatic Complexity metrics and Halstead Complexity metrics and design a system that will automatically measure software complexity of C-family programming languages.

3.0 METHODOLOGY

In this paper, the waterfall model of software engineering was adopted which is a linear sequential process that flows down with respect to its importance or an organized process to ensure that every stage is followed carefully before moving on to the next step. The steps taken were highlighted below:

(i) Requirement Engineering**:** This phase involved the gathering of required information online from different websites, eBooks and reviewing of existing software metrics applications.

(ii) System Design**:** The requirements gathered during phase one were the primary directing tool that was used in designing the system which measure complexity and defect rate of a code fragment.

(iii) System Coding: This phase deals with programming or development of the designed tool and creation of the database using C# programming language, .NET framework and Microsoft SQL server for database design.

(iv) System Testing: The system was tested efficiently by making sure no invalid data has been entered in forms in other not to disorganize the system, all pages were duly tested and the system was tested against exceptions and how they are handled. Black-box testing was possessed since this type of test is for the functionality of the system.

(v) Maintenance: The system maintenance was done on different versions as the system is being built while the code was maintained whenever new information is gathered based on the objects used on the system.

3.1 Selection of Metrics

The goal of this paper is to examine and implement a way in which the automatic measurement of source code complexity is possible to perform. One part of this paper is to find a suitable metric or a set of metrics that can indicate the complexity of a system. Therefore, the following software metrics have been studied and further implemented in developing the system.

Source line of code (SLOC): This metric is used to measure the quantitative characteristics of program source code. This metric is based on counting the lines of the source code. The limitations of SLOC may give an impression of a less useful metric, but wise use of this metric can still be valuable. Specifically, to the requirements of this paper, SLOC can be used to monitor change in terms of lines of code between different build versions of the software.

This metric can be calculated for individual functions of a program. If a function is too large in comparison with the average length per function, it may indicate that the function is hard to maintain and hence complex in terms of maintainability. The unit of this metric is a simple number that represents the number of lines in a source code file. This number can be used with other code metrics to formulate a new complexity indicator that can be more comprehensive.

LOC is usually represented as:

(i) kLOC: thousand lines of code

(ii) mLOC: million lines of code

Halstead software science: A suite of metrics was introduced by Maurice Howard Halstead in 1977. This suite of metrics is known as Halstead software science or as Halstead metrics. Halstead observed that metrics of software should reflect the implementation or expression of algorithms in different languages, but be independent of their execution on a specific platform. These metrics are therefore computed statically from the code. Most of the product metrics typically apply to only one particular aspect of a software product. In contrast, Halstead's set of metrics applies to several aspects of a program, as well as to overall production effort.

Halstead metrics are based on the following indices:

(i) n_1- distinct number of operators in a program

(ii) n_2 - distinct number of operands in a program

(iii) N_1- total number of operators in a program

(iv) N_2 - total number of operands in a program

Halstead formulas: Based on the above-mentioned indices (n_1, n_2, N_1, N_2), Halstead derived more than one formulas relating to the properties of program code. These formulas can measure Program Vocabulary (n), Program Length (N), Program Volume (V), Program Difficulty (D), Program Level (L), Total Effort (E), Development Time (T) and Number of Delivered Bugs (B). Halstead named his formulas as "Halstead's Software Science Metrics".

Program vocabulary $n = n_1 + n_2$ (1)

Program length $N = N_1 + N_2$ (2)

Program volume $V = N * log_2 n$ (3)

Difficulty $D = \frac{n_1}{2} \times \frac{N_2}{n_2}$ (4)

Program Level $L = (\frac{1}{D})$ (5)

Total effort $E = D \times V$ (6)

Development time $T = {}^{E}/_{S}$ (7)

The value of S is usually taken as 18 for these calculations.

Number of delivered bugs $B = \frac{E^{\frac{2}{3}}}{3000}$ (8)

McCabe cyclomatic complexity: Thomas McCabe (1976) introduced a metric based on the control flow structure of a program. This metric is known as McCabe cyclomatic complexity and it has been a famous code complexity metric throughout since it was first introduced. The McCabe metric is based on measuring the linearly independent path through a program and gives cyclomatic complexity of the program which is represented by a single number.

McCabe noted that a program consists of code chunks that execute according to the decision and control statements, e.g. if-else and loop statements. McCabe metric ignores the size of individual code chunks when calculating the code complexity but counts the number of decision and control statements. McCabe method maps a program to a directed, connected graph. The nodes of the graph represent decision or control statements. The edges indicate control paths that define the program flow. Cyclomatic complexity is calculated as:

$M = E - N + P$ (9)

where

M = McCabe metric,

E = the number of edges of the graph of a program,

N = the number of nodes of the graph and

P = the number of connected components.

P can also be considered as the number of exits from the program logic. The recommended ranges are shown in Table 1 below:

Table 1: McCabe cyclomatic complexity ranges

Cyclomatic Complexity	Code Complexity
1 – 10	A simple program, without much risk
11 – 20	More complex, moderate risk
21 – 50	Complex, high risk
50+	Untestable, very high risk

3.2 Requirement Engineering

This involved gathering necessary information in software engineering taking software metrics as a general topic to research on and having an overview of existing software metrics applications. The requirements gathered were then used in the system design to make the system user-friendly as good as possible. The system was developed with the aim of automatic measurement of source code complexity and increasing software reliability by measuring its defect rate. A defect rate is the percentage of output that fails to meet a quality target and it is calculated by testing output for non-compliance to a quality target (John, 2017). The following formula can be used to calculate defect rate;

$$Defect\ rate = \frac{Number\ of\ Delivered\ bugs}{Lines\ of\ code} \times 100 \quad (10)$$

Code defects are commonly measured as defects per thousand lines of code, which can be calculated with the following formula:

$$Defectsperthouandlinesofcode= \frac{NumberofDeliveredbugs}{1} x \frac{LInesofCode}{100} (11)$$

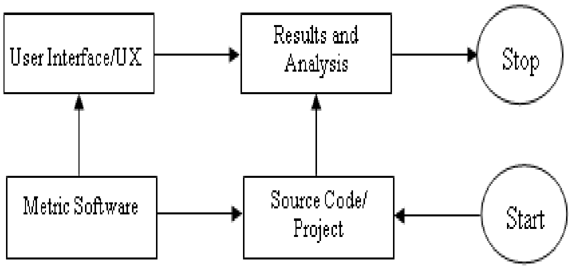


Figure 1: Block Diagram of Software Metric Application

The requirements needed to gather for every analysis carried out by the software being developed take a main requirement which was a source project or source code for another C# or other C-family programming languages where the software application runs some metrics and calculations based on the source code supplied. Figure 1 shows the block diagram for the system to work effectively and carry out a given task.

3.3 Software Testing

In improving the functionality and quality of the software, Black-box testing was adopted, which is a method of software testing that examines the functionality of an application without looking into its internal structures. This method of testing can be applied to virtually every level of software testing; unit testing, integration testing, system testing, and acceptance testing. However, the test is based on requirements and functionality, that is, on providing input, if the output matches the desired results, then the software is tested ok. Figure 2 shows the representation of black-box testing.

Furthermore, the software developers were given no specific instructions on how they should test the application, but they were encouraged to look for ways to break the application so that unanticipated bugs can be discovered and corrected. In this paper, net satisfaction index ranging from 0 to 100% was agreed with the software developers in the underlying test. The net satisfaction index rating scale maps ratings between completely dissatisfied and completely satisfied to numbers between 0 and 100% as seen in table 2.

Table 2: Net satisfaction index ranges

Net Satisfaction Index Rating	Remark
100%	Completely Satisfied
75%	Satisfied
50%	Neutral
25%	Dissatisfied
0%	Completely Dissatisfied

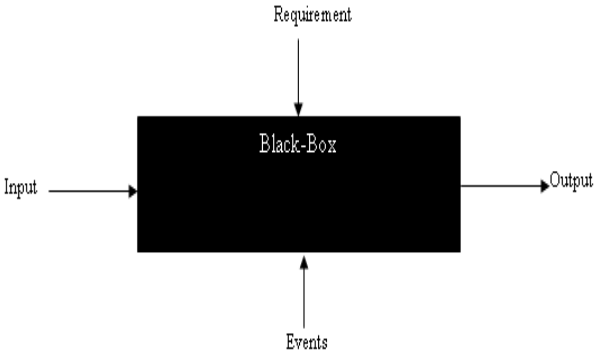


Figure 2: Black-Box Testing

4.0 RESULTS AND DISCUSSION

The selection of the measured source code was made after consulting with some software developers. Each software developers have different opinion about their source code before the complexity measurement. As stated, code-based complexity metrics Source Lines of Code (SLOC), Halstead Software Science and McCabe Cyclomatic Complexity were used in the development of this tool. However, the implementation of each complexity metrics are discussed below showing their results in the following figures.

4.1 Source Line of Code Metric (SLOC)

In order to evaluate Source Line of Code (SLOC) metric, this tool automatically counts and calculate the number of Blank Lines of Code (BLOC), Comment Lines of Code (CLOC) and Source Lines of Code (SLOC) which includes the actual code (logic and computation) and declarations in the source code while BLOC and CLOC improves readability and helps in understanding the code as well as during maintenance. SLOC metric is used to measure the quantitative characteristics of the program source code and gives the software size estimations. However, in increasing the software reliability the defect rate was measured with the calculated number of delivered bugs and lines of code. Table 3 shows the summary of the SLOC metric of the measured program source code.

Table 3: Summary of the SLOC Metric

Metric Attributes	Programming Languages				
	C++	C	Python	C#	JavaScript
BLOC	4	314	291	245	316
CLOC	0	315	210	156	293
SLOC	82	2005	1807	1162	1828
Total LOC	86	2634	2308	1563	2437
Defect Rate	0.327	1.636	73.910	56.793	1.349

4.2 Halstead metrics

To evaluate Halstead metrics, this tool counts the distinct number of operators, distinct number of operands, total number of operators and total number of operands in a program source code. In order to achieve this, considering a C program. Examples of unique operators are main, ( ), { }, int, scanf, &, =, +, /, printf and so on. Also, examples of operands are a, b, c, 3, avg, %d e.t.c. The tool finds all the distinct operators and operands in the source code and counts how many times each distinct operator or operands appears in the source, then total them and gives the output as total number of operators and total number of operands. Once these values have been computed then the Halstead set of metrics can be calculated as stated in each formula discussed in the previous chapter. The accuracy depends on the tool evaluating every possible places in the model that could have a predefined operator.

However, SLOC metric is typically applied to only one particular aspect of a software product. In contrast, Halstead set of metrics applies to several aspect of a program source code as it is good to have more than one metric for the quantitative measure of a program and not to fully rely on one. Although, Halstead metrics are difficult to compute manually as it is not easy to count the number of operators and operands if a program is using large number of operators and operands. Hence, in this research automatic measurement of source code is achieved by developing a tool that solves the limitations of the existing tool. Table 4 shows the summary of the measured source code.

4.3 McCabe cyclomatic complexity

McCabe is a measure of the complexity of a module's decision structure. For the McCabe metric, this tool searches for IF, FOR, WHILE, CASE statements in the source code and counts them to calculate the McCabe complexity value. It counts the number of decision or control structure and maps a program to a directed, connected graph. The nodes of the graph represent decision or control statements and the edges indicate control path that define the program flow. However, since no organization or software company will want to release their source code and because of the license time restrictions two senior software developers were consulted. The selected source code for performing the code analysis is Python and C#.

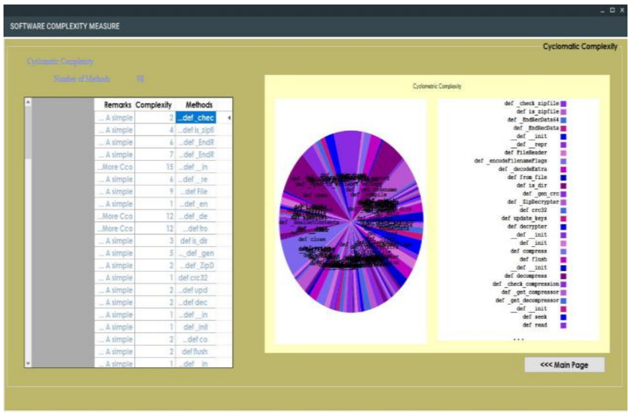
The software developer's opinion about Python was that the package is well structured and designed which should possibly contain less complex code. In other words, the software package C# was supposed to be more complex as said by the software developer. Hence, the McCabe cyclomatic complexity metric used in this tool results for a fast assessment and from the calculated results the Python software package has less complex code while the C# software package has more complex code which relates with the software developer's opinion. Figure 3 shows the calculated result of the Python software package and Figure 4 shows the calculated result of the C# software package.

4.4 Evaluation of the Tool

The developed tool has a view in which the effectiveness of waterfall model is being implemented. The outcome of the developed tool shows that it has the ability of performing automatic measurement of source code on regular basis and also extend the functionality of the existing tool. However, black-box testing was adopted to test the tool based on requirements and functionality by the software developers because it is a software testing method which examines the functionality of an application without looking into its internal structure or workings that is, in the testing method, the design and structure of code are not known to the software tester and testing engineers and end users conduct this test on software. The software developers were chosen in the software testing stage and the test was conducted on the system.

Table 4: Summary of the Halstead Metrics

Metric Attributes	**Programming Languages**				
	C++	C	Python	C#	JavaScript
Distinct Operators n_1	29	32	81	121	25
Distinct Operands n_2	30	37	1330	585	24
Total Operators N_1	76	217	7460	6860	168
Total Operands N_2	59	131	8421	4854	122
Program Length N	135	348	15881	11714	290
Program Vocabulary n	59	69	1411	706	49
Program Volume V	794.157	2125.767	166154.999	110855.725	1628.266
Program Difficulty D	28.517	56.649	256.429	501.995	63.542
Program Level L	0.035	0.018	0.004	0.002	0.016
Total Effort E	22646.975	120422.575	42606960.240	55645915.710	103463.278
Development Time T	1258.165 Sec.	6690.143 Sec.	2367053.347 Sec.	3091439.762 Sec.	5747.960 Sec.
Delivered Bugs B	0.268	0.816	40.902	48.875	0.738



SOFTWARE COMPLEXITY MEASURE

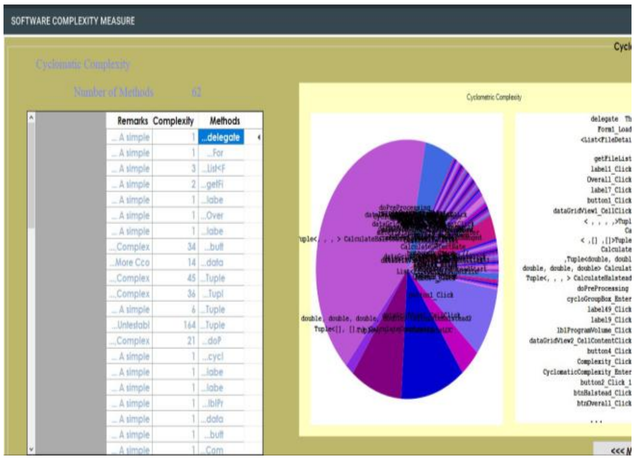
Cyclomatic Complexity

Cyclomatic Complexity

Number of Methods 98

Remarks	Complexity	Methods
... A simple	2	...def _chec
... A simple	4	...def is_zipfi
... A simple	6	...def _EndR
... A simple	7	...def _EndR
...More Cco	15	...def __in
... A simple	6	...def __re
... A simple	9	...def File
... A simple	1	...def _en
...More Cco	12	...def _de
...More Cco	12	...def fro
... A simple	3	def is_dir
... A simple	5	..._def _gen
... A simple	2	...def _ZipD
... A simple	1	def crc32
... A simple	2	...def upd
... A simple	2	...def dec
... A simple	1	...def __in
... A simple	1	def _init
... A simple	2	...def co
... A simple	2	def flush
... A simple	1	...def __in

Figure 3: Desktop Interface Showing Python Software Package Result



SOFTWARE COMPLEXITY MEASURE

Cycl

Cyclomatic Complexity

Number of Methods 62

Remarks	Complexity	Methods
... A simple	1	...delegate
... A simple	1	...For
... A simple	3	...List<F
... A simple	2	...getFi
... A simple	1	...labe
... A simple	1	...Over
... A simple	1	...labe
...Complex	34	...butt
...More Cco	14	...data
...Complex	45	...Tuple
...Complex	36	...Tupl
... A simple	6	...Tuple
...Untestabl	164	...Tuple
...Complex	21	...doP
... A simple	1	...cycl
... A simple	1	...labe
... A simple	1	...labe
... A simple	1	...lblPr
... A simple	1	...data
... A simple	1	...butt
... A simple	1	...Com

Figure 4: Desktop Interface Showing C# Software Package Result

The ratings are clearly stated in table 5 showing the software developer’s opinion about the developed tool. Hence, this method attempts to find errors in functionality and quality of the system as shown in figure 5. The overall average for functionality and quality of the system results to 90% and 90% respectively. Hence, it is obvious that on providing the input by the end users, the output matches with the desired results and the developed tool is said to be tested ok.

Table 5: Summary of Black-box Testing for each Software Developer

Programming Languages	Functionality	Quality
C++	100%	100%
C	100%	100%
Python	75%	75%
C#	75%	75%
JavaScript	100%	100%
Average	90%	90%

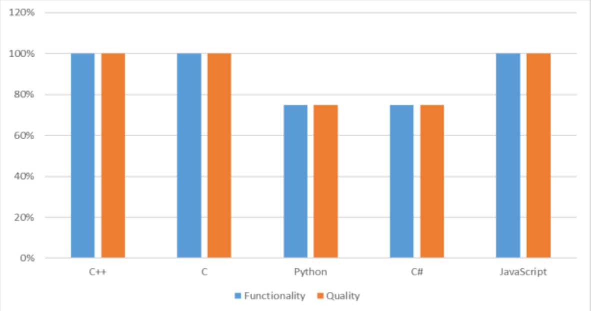


Figure 5: Summary of Black-box Testing for each Software Developer

5.0 CONCLUSION

As the demand for software is growing at an exponential rate, the complexity of software is also increasing. The program complexity measure seems to be the most capable measure for both quantitative and control flow of the software project. In order to understand and measure the quality and functionality of a software, it is important to relate it to measurable quantities. However, identifying a reason for gathering metrics is essential to ensure relevant metrics are gathered because too many metrics become confusing if they are not required.

In this paper, code-based metrics Source Line of Code (SLOC), McCabe Cyclomatic Complexity and Halstead Complexity Metrics were studied and used to achieve the aim of developing an automated software complexity measurement tool. The metrics provide visibility and control for the complex software development process and therefore, they are valuable tools for providing guidance on improving the software development process as well as meeting organizational goals to improve software productivity and quality. The development of this tool has been built successfully and tested efficiently to run on Microsoft Windows operating system. However, software testing is an important phase of the software development lifecycle. It is important to optimize test cases and generate them automatically so as to minimize testing time, effort and cost. Hence, black-box testing was adopted in this paper as it focuses on examining the functionality of the system.

REFERENCES

Bhatti, H. R. (2010). Automatic Measurement of Source Code Complexity. Master’s Thesis, Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology. pp 1-14.

Binstock, A. (2010). Integration Watch: Using Metrics effectively. (http://www.sdtimes.com/link/34157). SD Times. BZ Media.

Borchert, T. (2008). Code Profiling: Static Code Analysis. Master’s Thesis, Department of Computer Science, Karlstad University, Sweden. pp 12-16

Chandra, E., Linda, P. and Edith, A. (2010). Class Break Point Determination Using CK Metrics Thresholds, *Global Journal of Computer Science and Technology*, Vol.10, Issue.14, Pp:73-77.

Chidamber, S. R. and Kemerer, C. F. (1994). A metrics suite for object-oriented design. *IEEE Transactions on Software Engineering*. Vol 20 (6). Pp. 476-498.

Fenton, N. and Bieman, J. (2014). Software metrics. A Rigorous and Practical Approach. Third Edition. CRC Press. pp 1 – 67.

Halstead, M. (1977). Elements of Software Science, Operating and Programming Systems Series, Elservier Computer Science Library North Holland, N.Y. Elsevier North-Holland, Inc. I SBN 0-444-00205-7. pp 1 – 6.

Jay, G., Hale, J. E., Smith, R. K., Hale, D., Kraf, N. A. and Ward, C. (2009). Cyclomatic complexity and lines of code: Empirical evidence of a stable linear relationship, *J. Software Engineering & Applications*, pp. 7.

Jian Li, Pinkie He, Jieming Zhu and Micheal Lyu (2017). Software Defect Prediction via Convolutional Neural Network. Department of Computer Science and Engineering. The Chinese University of Hong Kong, China. *IEE International Conference on Software Quality, Reliability and Security*. Pages 8 – 10.

John Spacey (2017). How Defect Rate is Calculated. https://simplicable.com/new/defect-rate.

Kaner Cem (2004). Software Engineer Metrics. What do they measure and how do we know? (http: //citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.2542).

Lincke, R., Lundberg, J. and Löwe, W. (2008). Comparing software metrics tools. *In Proceedings of the 2008 International Symposium on Software Testing and Analysis. ISSTA'08*. pp. 131–142. New York, NY:ACM.

McCabe, T. J. (1976). A Software Complexity Measure. *IEEE Transactions on Software Engineering*. Vol. SE-2, No. 4, pp. 308-320.

Olabiyisi, S. O., Omidiora, E. O. and Sotonwa, K. A. (2013). Comparative Analysis of Software Complexity of Searching Algorithms Using Code Based Metrics. *International Journal of Scientific and Engineering Research*. Vol. 4, Issue 6. ISSN: 2229-5518. pp 1-2

Panovski, G. (2008). Product Software Quality. Master's Thesis, Department of Mathematics and Computing Science, Technische Universiteit, Eindhoven. pp 1-24.

Sanjeev, A. and Barak, B. (2009). Computational Complexity. A Modern Approach, Cambridge. ISBN 978-0-521-42426-4. Zbl 1193.68112. page 1- 8.

Silva, D., Koadagoda, N. and Perera, H. (2012). Applicability of three complexity metrics, in Proc. *International Conference on Advances in ICT for Emerging Regions*. pp 12-16.

Silva, D., Weerawarna, N., Kuruppu, K., Ellepola, N. and Kodagoda, N. (2013). Applicability of three cognitive complexity metrics. *In Proc. 2013 8th International Conference on Computer Science & Education*. pp 9 -12.

Scotto M., Sillitti, A., Succi, G. and Vernazza, T. (2004). A Relational Approach to Software Metrics. *ACM Symposium on Applied Computing*. pp 1536 - 1540.

Widheden, K and Göran, J (2010). Software Complexity: Measures and Measuring for Dependable Computer Systems ", MASTER'S THESIS, Department of Computer Science and Engineering, University of Gothenburg, Goteborg, Sweden. Page 1 – 14.

Yu, T., Wen, W., Han, X. and Hayes, J. (2018). Conpredictor: Concurrency Defect Prediction in Real-World Applications. *In IEEE International Conference on Software Testing, Verification and Validation*. Pages 168-179.

Zhang, M. and Baddoo, N. (2007). Performance comparison of software complexity metrics in an open source project, *In Proc. 14th European Conference Software Process Improvement*, Potsdam, Germany. Page 1 – 16.

ESTABLISHING CONSENSUS IN KNOWLEDGE BASE CREATION OF MEDICINAL PLANTS OF AFRICAN TRADITIONAL MEDICINE PRACTICE

[1*]Awoniran O. M., [2]Soriyan H. A., [3]Elujoba A. A. and [4]Okunnade A. A.

[1]Department of Computer Science & Information Technology, Bowen University, Iwo, Nigeria
[2]Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
[3]Department of Pharmacognosy, Obafemi Awolowo University, Ile-Ife, Nigeria
[4]Department of Physics & Engineering Physics, Obafemi Awolowo University, Ile-Ife, Nigeria

*Corresponding author's e-mail: *lekanawoniran@gmail.com*

ABSTRACT

This research work creates a harmonized and holistic knowledge base of medicinal plants in African traditional medicine practice. The methodology employed in the research work involves three stages of the Delphi knowledge elicitation technique conducted for nine (9) experts in the ATM practice. Knowledge collation and elimination were carried out after each stage of the knowledge elicitation process till a perfect consensus and agreement was achieved. These were both measured using Kendall's coefficient of concordance (w) and the Fleiss' kappa agreement analysis (k). Results from the work show a progression in the consensus and agreements on the medicinal plants and herbs useful for the treatment and management of some diseases in the ATM practice. At the final stage of the interview, perfect consensus and agreement were achieved. In conclusion, the use of the Delphi technique as a knowledge elicitation tool aids consensus and agreement in the creation of a domain knowledge base. This will aid the development of a consensual, holistic and non-subjective knowledge base of medicinal plants in the African traditional medicine practice.

Keywords: Knowledge Based Systems, Knowledge Elicitation, Delphi Knowledge Elicitation Technique, African Traditional Medicine Practice

1.0 INTRODUCTION

A knowledge base is a set of underlying facts, assumptions, and rules in a knowledge domain [1]. These facts, assumptions or rules range from simple assertions to complex procedures about entities that may really or fundamentally exist in the knowledge domain. The creation of a knowledge base usually requires a process of knowledge elicitation from experts in the domain [2]. During the process, domain experts provide knowledge about various aspects of the knowledge domain. Various methods are in existence for the purpose of knowledge elicitation. These include interviews, observation, storytelling, round-table among others [3]. The results of an elicitation process are usually qualitative in nature [4]. In situations whereby there is more than one expert from which knowledge was elicited, establishing a consensus from the qualitative data obtained then becomes a task; otherwise, the development of a holistic knowledge base would be near impossible or at best, subjective. This problem is more complex in knowledge domains with unstructured data [5].

Unstructured data is a type of data that cannot be easily represented using tables [6]. This is because they lack a definite model and are not organized in a pre-defined manner [7]. Data and information in most knowledge domains are unstructured in nature [8]. This makes the process of knowledge elicitation and ultimately the creation of knowledge bases in these domains quite difficult. Such is the case in African traditional medicine (ATM) practice.

In the ATM practice, therapeutic, processing, preparatory and usage knowledge are unstructured in nature [9]. Therefore, the establishment of a consensus among knowledge donors in the practice is important for the creation of a holistic knowledge base.

Delphi technique provides a means of achieving consensus among knowledge donors [10]. The technique aims at acquiring the most reliable consensus of opinion by a group of experts through a series of intensive interviews along with an opinion control feedback [11].

This research work aims at achieving consensus in the creation of a knowledge base of medicinal plants and herbs in the ATM practice. It focuses on the use of medicinal plants and herbs for the treatment and/or management of various diseases and ailments in the African traditional medicine (ATM) practice.

This paper is structured as follows; Section 1.0 gives an introduction and background to the research, Section 2.0 reviews existing works related to the knowledge elicitation and knowledge base creation. Section 3.0 discusses the research methodology the methods employed in carrying out the research work. Section 4.0 shows the results obtained from the work and discusses the implications of the results. Finally, section 5.0 concludes the research work.

2.0 LITERATURE REVIEW

Knowledge can be defined as the practical or theoretical understanding of facts, descriptions or principles of a particular subject [12]. These facts, descriptions or principles are usually gathered by education or experience. Knowledge can be implicit or explicit in nature. Implicit knowledge is also regarded as tacit knowledge and it is the type of knowledge that basically resides in the mind of the knowledge owner [13]. This type of knowledge is hard to articulate and knowledge owners usually cannot describe it other than their own skillful display. Explicit knowledge, on the other hand, is the knowledge that has been codified and documented using some sort of media [13].

Most expert-driven knowledge is implicit in nature and hence difficult to codify. Also, expert-driven knowledge is a function of the perceptions and experiences of the individual experts who possess this knowledge [14]. These facts are reasons why knowledge is regarded as a complex entity and also, the reason why knowledge elicitation from experts is more than the simple process of data acquisition. Methods that have been employed in the past for the purpose of knowledge elicitation include, interviews, storytelling, laddering, protocol analysis, critical decision method, concepts sorting, etc [15].

Knowledge elicited from experts is most probably likely to be represented as a subjective distribution [16]. These distributions play an important role in inference and decision-making in the development of knowledge models [17]. Methods through which knowledge elicitation can be made scientific in order to forestall subjectivity was proposed using protocols [18]. The work argues that by following a recognized protocol that is designed to address and minimize cognitive biases that experts are prone to when making judgments, subjectivity in such judgments will likely be eliminated.

Again, the consensus of elicited knowledge could also be achieved using information integration [19]. Cognitive task analysis for information triangulation was proposed as a method for information integration of the needs, goals, and design requirements for the design of a newborn intensive care unit [20]. The method involves a representative definition of design goals by accounting for the perspectives of all user groups involved with the unit. Results from the work show that each role has a unique perspective and a specific detail based on the role goals. The work concludes that source triangulation should be employed as part of methods to achieve design objectives in complex socio-technical systems.

Also, the Delphi technique was used for establishing consensus on core competencies of the model of community and family medicine department in India [21]. In the study, experts were contacted by e-mail and their feedbacks were analyzed. Results from the work show consensus being reached on various core competencies in the department which are “Clinician”, “Primary-Care Physician”, and “Professionalism”. However, the work claimed that more experienced experts were less consistent in their responses.

Furthermore, a modified Delphi technique was employed to inform the development of best practices in interprofessional training for collaborative primary healthcare [22]. In the study, a modified Delphi technique was used in knowledge elicitation using three rounds of interviews. The first round of interviews was carried out with the use of questionnaires. The qualitative results obtained was then written out in a Likert scale-like version and presented to the respondents in the second round of interview. Finally, the results were collated and re-presented to the respondents in the final round of interviews. The work demonstrated a successful elicitation of knowledge from a heterogeneous panel of stakeholders in primary health education and training [23].

Specifically, in ATM practice, a framework for knowledge capture was proposed to enhance the codification and preservation of knowledge in ATM treatment of Malaria [24]. The methodology adopted in the work involves the elicitation of knowledge by means of interviews from ATM practitioners. The knowledge derived was structured and then built into a knowledge-based system (KBS). The KBS was tested along with experts in the field and a high level of agreement was recorded between the KBS and the experts.

Finally, research on factors influencing indigenous knowledge data elicitation from herbal medicinal practitioners was conducted in south-western Nigeria [25]. The work highlights the fact that the lack of or inadequate documentation of indigenous knowledge contributes majorly to the extinction (or the threat of it) of such knowledge. Methods from the work include interviews and direct observation of processes in traditional medicine. Results from the work show that familiarity and compensation are key factors that facilitate openness of knowledge holders to share information.

3.0 METHODOLOGY

The methodology employed in the research work involves firstly, the selection and recruitment of ATM practitioners into the study, thereafter, the initial interview with designed questionnaire, knowledge collation, agreements analysis, and follow-up interviews was conducted. A diagram showing the processes involved in the methodology is provided in figure 1.1.

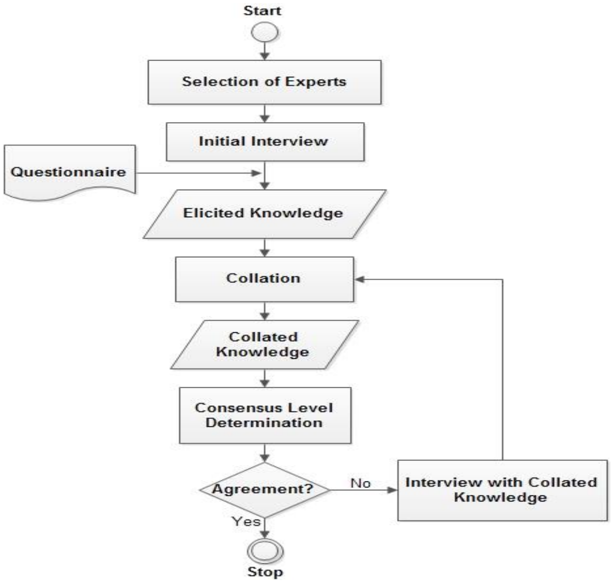


Figure 1.1: The Methodological Framework

3.1 Selection of Experts

In the selection and recruitment of ATM practitioners for the study, four criteria were considered in order to recruit bona fide ATM practitioners. The first criterion that was considered in recruiting the ATM practitioners was the coverage area [26, 27]. This study considered ATM practitioners from the Southwest of Nigeria only. This is to enhance precision and improve the specificity of the expected results. Also, the availability of medicinal plants is subject to geographical dispositions, hence achieving consensus in multiple geographical regions might be impossible.

The second criterion considered in the selection of ATM practitioners for the study was the mode of training as a practitioner. For the study, ATM practitioners that were trained due to ancestral lineage or those who have gone through a master-servant type of apprenticeship were recruited into the study. This is because the practitioners that practice the trade due to ancestral lineage are usually more trained and experienced. Also, the master-servant type of apprenticeship provides trainees with more hands-on experience than other types of training methods. The third criterion was the disease specialties and nature of the practice. ATM practitioners specialize in more than one particular type of disease [28]. However, some ATM practitioners have more knowledge of some particular diseases than they do on others. For the study, only practitioners in full-time practice and whose specialties are in the disease areas that have been known over the years to be successfully treated with traditional medicines were considered.

The last consideration used in the selection of the ATM practitioners was the knowledge application. This relates to knowledge utilization by the practitioners. Under this criterion, conditions such as the number of patients treated per week, the number of patients successfully managed and certified cured out of the number treated in a week were considered. For the study, ATM practitioners that consult patients at least 10 patients per week with at least 5 been successfully treated were recruited for the study.

3.2 Initial Interview

The initial interview was conducted with the use of an interview guide. The various aspects of the interview are expertise in traditional medicine practice, knowledge acquisition in traditional medicine practice, knowledge impact in traditional medicine practice, the practice of African traditional medicine (diseases and treatments), the practice of African traditional medicine (pharmaceutical aspects). The interview questions were designed to enable discussions on the various aspects of the interview. The language of communication of the interviews is the *Yoruba* language. After the interviews, language transcription was then carried out to convert the results obtained to English language.

3.3 Knowledge Collation

After the initial interviews were conducted, the various responses gathered from the respondents were transcribed to English language. The various medicinal plants and herbs useful for the treatments of various diseases and ailments as given by the respondents were then collated. Concordance and agreements analysis was then conducted in order to determine the level of consensus in the knowledge elicited.

3.4 Consensus Analysis

Upon the successful collation of data, the level of consensus in the elicited knowledge was then determined by analyzing concordance and agreements in the elicited knowledge. The level of concordance was determined using Kendall's coefficient of concordance (w) and the level of agreements was determined using the Fleiss' kappa agreement analysis (k). Both Kendall's coefficient of concordance (w) and the Fleiss' kappa agreement (k) have values that range from 0.0 to 1.0 [28, 29]. For Kendall's coefficient of concordance (w), 0.0 denotes minimum concordance while 1.0 denotes maximum concordance. For Fleiss' kappa agreement analysis (k), 0.0 denotes minimum agreement while 1.0 denotes maximum agreement. However, for both cases, 8.0 is considered as an almost perfect concordance or an almost perfect agreement. The equation for Kendall's coefficient of concordance (w) is given below where r_{ij} is the number of medicinal plants given by practitioner *j* for ailment *i*, R_i is the total number of medicinal plants selected for a particular illness and $\bar{R}$ is the mean of the medicinal plants' selections. Furthermore, *m* is the number of respondents, *n* is the number of ailment for which medicinal plants are requested and *S* is the sum of squared deviations [30].

$$R_i = \sum_{j=1}^{m} r_{ij} \qquad \textbf{eqn 1}$$

$$\bar{R} = \frac{m(n+1)}{2} \qquad \textbf{eqn 2}$$

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2 \qquad \textbf{eqn 3}$$

Kendall's coefficient of concordance (w) =

$$\frac{12S}{m^2(n^3 - n)} \qquad \textbf{eqn 4}$$

The equation for Fleiss' kappa agreement analysis (k) is given below where $\bar{p}$ denotes the extent to which respondents agree for the treatment/management of a particular disease/ailment and $\bar{p_e}$ denotes the agreements for herbs used in treating a particular disease/ailment.

Fleiss' kappa agreement analysis (k) =

$$\frac{\bar{p} - \bar{p_e}}{1 - \bar{p_e}} \qquad \textbf{eqn 5}$$

3.5 Follow-Up Interviews

After the consensus level determination, the practitioners were asked to select useful medicinal plants and herbs for the treatment /management of the various diseases elicited previously from the collated list of medicinal plants and herbs. After this, knowledge collation and consensus level determination were once again carried out. This process was repeated until perfect concordance and agreements were obtained.

4.0 RESULTS AND DISCUSSION

A total of nine (9) experts were recruited into the study taking into consideration the criteria for the selection of experts outlined in section 3.0. After the initial interview was carried out, knowledge of twenty-two diseases and ailments were elicited. These are Asthma *(Ikoo Semisemi)*, Chicken Pox *(Arun Saponna)*, Cold & Catarrhal *(Otutu/Osin)*, Convulsions *(Giri)*, Cough *(Iko)*, Diabetes *(Ito Suga)*, Diarrhea *(Igbe Gbuuru)*, Eczema *(Ifo)*, Fibroid *(Iju)*, Gonorrhea *(Atosi)*, Hernia *(Ipa)*, Hypertension *(Eje Riru)*, Erectile Dysfunction *(Idakole)*, Malaria *(Iba)*, Pile *(Jedi-Jedi)*, Rheumatism *(Aromolapa/Aromomlegun)*, Ring Worm *(Lapalapa)*, Stomach Disorder (*Inu rirun*), Syphilis *(Atosi Ako)*, Typhoid *(Iba Jefun/Jedo)*, Ulcer *(Ogbe inu)*, and Yellow Fever *(Iba Aponju-ponto)*. The diseases and ailments are labeled D_1 to D_{22} respectively.

Furthermore, at the stage, ten medicinal plants and herbs were obtained for D_1, seven for D_2, nine for D_3, thirteen for D_4, nine for D_5, nine for D_6, ten for D_7, three for D_8, three for D_9, nine for Gonorrhea D_{10}, five for D_{11}, six for D_{12}, nine for D_{13}, fifteen for D_{14}, Six for D_{15}, nine for D_{16}, three for D_{17}, nine for D_{18}, eleven for D_{19}, thirteen for D_{20}, seven for D_{21}, and twelve for D_{22}. This is presented in Table 1.1

In the second round of interviews, the practitioners were asked to select the medicinal plants and herbs useful in the treatment/management of the various diseases from the list of medicinal plants and herbs obtained at the first round of interviews. Table 1.2 shows the result of the second round of interviews. Afterward, the collation of the medicinal plants and herbs was carried out.

Again, in the third round of interviews, the practitioners were asked to select medicinal plants and herbs useful in the treatment/management of the various diseases from the list of medicinal plants and herbs obtained at the second round of interviews. Table 1.3 shows the result of the third round of interviews. In the third round of interviews, all respondents agree that all the medicinal plants and herbs presented for the treatment of various diseases/ailments were useful in the treatment of the diseases/ailments.

The results of Kendall's coefficient of Concordance (W) and the Fleiss' kappa agreements (k) analysis for the various interview rounds are given in table 1.4 and figure 1.2 respectively. Finally, the consensual knowledge base of medicinal plants and herbs of ATM practice after the final round of the Delphi process is given in table 1.5.

From the results obtained, Kendall's coefficient of concordance shows a value of 0.27 (Poor concordance) at the first round of the Delphi process, 0.79 (Near perfect concordance) at the second round of the process and finally 1.00 (Perfect concordance) at the final round of the process. This shows a positive progression in the establishment of the consensus of the medicinal herbs used for the treatment/management of the various diseases. The same can also be said for the Fleiss kappa agreements analysis, however, at a more detailed level of the individual diseases/ailments. For all the diseases/ailments, there was a progression from a very low agreement to a very high agreement. Achieving perfect consensus and agreement are important in the research since the actual medicinal plants useful in the treatment of the ailments are required.

Table 1.1: Number of Herbs obtained from Respondents after the Initial Interview

D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}	D_{18}	D_{19}	D_{20}	D_{21}	D_{22}
10	7	9	13	9	9	10	3	3	9	5	6	9	15	6	9	3	9	11	13	7	12

Table 1.2: Number of Herbs Selected by Respondents after the Second Round of Interview

Resp	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}	D_{18}	D_{19}	D_{20}	D_{21}	D_{22}
1	7	5	7	10	6	7	9	2	2	7	4	6	7	14	5	7	2	7	10	11	5	10
2	8	4	6	11	6	6	9	2	2	8	3	5	8	13	4	7	2	7	9	12	6	10
3	7	4	7	11	5	6	8	2	2	7	3	6	7	13	3	8	2	7	9	11	5	10
4	6	3	7	11	6	6	9	2	2	7	4	5	7	10	4	7	2	7	9	12	5	11
5	7	4	7	10	4	6	9	2	2	7	3	5	7	13	4	6	2	6	9	11	5	10
6	7	4	6	11	5	7	9	2	2	7	2	5	7	11	4	7	2	7	8	10	6	10
7	6	3	7	11	5	6	9	2	2	6	3	5	7	13	4	5	2	7	9	11	5	9
8	7	4	6	10	6	6	8	2	2	7	3	5	6	10	3	7	2	6	8	11	5	10
9	7	3	7	11	6	5	9	2	2	6	3	5	5	11	2	7	2	5	7	10	4	9

Table 1.3: Number of Herbs Selected by Respondents after the Third Round of Interview

Resp	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}	D_{18}	D_{19}	D_{20}	D_{21}	D_{22}
1	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
2	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
3	7	4	7	11	5	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
4	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
5	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
6	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
7	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
8	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10
9	7	4	7	11	6	6	7	2	2	7	3	5	7	13	4	7	2	7	9	11	5	10

Table 1.4: Kendall's Coefficient of Concordance (W) Results

Round	m	n	$\bar{R}$	*S*	W
1	9	22	103.5	22774	0.27
2	9	22	103.5	56859	0.79
3	9	22	103.5	89803	1.00

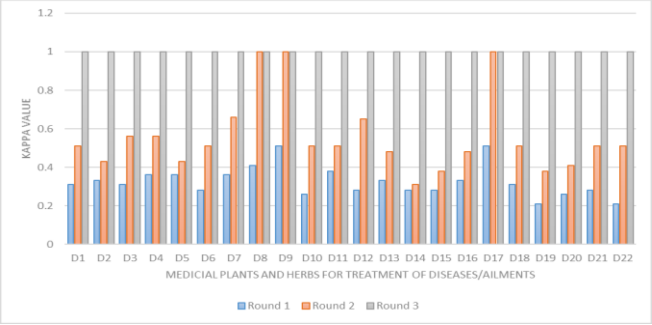


Table 1.5: Elicited Knowledge Base of Medicinal Plants and Herbs of ATM Practice after the Final Round of the Delphi Process

S/No.	Disease/ Ailment	Medicinal Plants for Treatment after the First Round of Interview	Medicinal Plants for Treatment after the Second Round of Interview	Medicinal Plants for Treatment after the Third Round of Interview
1	Asthma	Baobab *(Igi-Ose)*, Garlic *(Ayuu)*, Giant Milk Weed *(Bomubomu)*, Asthma plant *(Oroigi)*, Indian shot *(Idodo)*, Silk Cotton Tree *(Araba)*, Clausena *(Oboku)*, Desmodium *(Epa-Ile)*, Starburr *(Dangunro-Gogoro)*, Heliotrope *(Agogo-Igun).*	Baobab *(Igi-Ose)*, Garlic *(Ayuu)*, Giant Milk Weed *(Bomubomu)*, Asthma plant *(Oroigi)*, Indian shot *(Idodo)*, Silk Cotton Tree *(Araba)*, Clausena *(Oboku)*, Desmodium *(Epa-Ile).*	Baobab *(Igi-Ose)*, Garlic *(Ayuu)*, Giant Milk Weed *(Bomubomu)*, Asthma plant *(Oroigi)*, Indian shot *(Idodo)* Silk Cotton tree *(Araba)*, Desmodium *(Epa-Ile)*.
2	Chicken Pox	Alligator Pepper *(Ata-Ire)*, Tallow Tree *(Orogbo Erin)*, Sasswood *(Olu-Obo)*, Sesame *(Eluru)*, *Cayenne Pepper (Bawa)*, Tea bush *(Efinrin-Nla)*, Tea bush *(Efinrin-Nla)*.	Alligator Pepper *(Ata-Ire)*, Tallow Tree *(Orogbo Erin)*, Sasswood *(Olu-Obo)*, Sesame *(Eluru)*, *Cayenne Pepper (Bawa).*	Alligator Pepper *(Ata-Ire)*, Tallow Tree *(Orogbo Erin)*, Sasswood *(Olu-Obo)*, Sesame *(Eluru)*.
3	Cold & Catarrhal	Guava *(Gilofa)*, *Garlic (Ayuu)*, *Lemon Juice (Ọsàn –Lakuregbe)*, Flacourtia *(Osere)*, Moringa Tree *(Ewe-Ile)*, Sweet and Hairy Basil *(Efinrin-Wewe*), *Cayenne Pepper (Bawa),* Starburr *(Dangunro-Gogoro)*, Ginger *(Ata-Ile)*.	Guava *(Gilofa)*, *Garlic (Ayuu)*, *Lemon Juice (Ọsàn –Lakuregbe)*, Flacourtia *(Osere)*, Moringa Tree *(Ewe-Ile)*, Sweet and Hairy Basil *(Efinrin-Wewe*), *Cayenne Pepper (Bawa),* Starburr *(Dangunro-Gogoro)*.	Guava *(Gilofa)*, *Garlic (Ayuu)*, *Lemon Juice (Ọsàn –Lakuregbe)*, Flacourtia *(Osere)*, Moringa Tree *(Ewe-Ile)*, Sweet and Hairy Basil *(Efinrin-Wewe*), *Cayenne Pepper (Bawa).*
4	Convulsions	Cascarilla *(Eru)*, Desmodium *(Epa-ile)*, Cotton *(Owu)*, Heliotrope *(Agogo-Igun),* Flacourtia *(Osere)*, Moringa Tree *(Ewe-ile)*, Pergularia *(Atufa)*, Guava *(Gilofa),* Sweet basil *(Efinrin-wewe)*, Secamore *(Arilu)*, Hog plum *(Akikan),* Latex *(Abara)*, Adenia *(Arokeke)*.	Cascarilla *(Eru)*, Desmodium *(Epa-ile)*, Cutton *(Owu)*, Heliotrope *(Agogo-Igun),* Flacourtia *(Osere)*, Moringa Tree *(Ewe-ile)*, Pergularia *(Atufa)*, Guava *(Gilofa),* Sweet basil *(Efinrin-wewe)*, Secamore *(Arilu)*, Hog plum *(Akikan),* Latex *(Abara)*.	Cascarilla *(Eru)*, Desmodium *(Epa-ile)*, Cutton *(Owu)*, Heliotrope *(Agogo-Igun),* Flacourtia *(Osere)*, Moringa Tree *(Ewe-ile)*, Pergularia *(Atufa)*, Guava *(Gilofa),* Sweet basil *(Efinrin-wewe)*, Secamore *(Arilu)*, Hog plum *(Akikan).*
5	Cough	Crab's eye *(Oju-Ologbo)*, Starburr *(Dangunro-Gogoro)*, Bear's breech *(Ahon-Ekun)*, Adenia *(Arokeke)*, Ginger *(Ata-Ile)*, Lemon Juice *(Ọsàn -Lakuregbe),* Alligator Pepper *(Ata-Ire)*, Tea bush *(Efinrin-Nla)*, Bitter leaf *(Ewuro)*.	Crab's eye *(Oju-Ologbo)*, Starburr *(Dangunro-Gogoro)*, Bear's breech *(Ahon-Ekun)*, Adenia *(Arokeke)*, Ginger *(Ata-Ile)*, Lemon Juice *(Ọsàn -Lakuregbe),* Alligator Pepper *(Ata-Ire)*.	Crab's eye *(Oju-Ologbo)*, Starburr *(Dangunro-Gogoro)*, Bear's breech *(Ahon-Ekun)*, Adenia *(Arokeke)*, Ginger *(Ata-Ile)*, Lemon Juice *(Ọsàn -Lakuregbe).*
6	Diabetes	Stool wood *(Ahun)*, Seeds of Hunteria *(Abere),* Bitter leaves *(Ewuro)*, Moringa *(Ewe-Ile)*, Scent Leaf *(Efinrin)*, Dutchman's Pipe *(Ako-Igun)*, Bitter leaf *(Ewuro)*, Wonderful kola *(Obi-Ata)*.	Stool wood *(Ahun)*, Seeds of Hunteria *(Abere)*, Bitter leaves *(Ewuro)*, Moringa *(Ewe-Ile)*, Scent Leaf *(Efinrin)*, Dutchman's Pipe *(Ako-Igun)*, Bitter Leaf *(Ewuro)*.	Stool wood *(Ahun)*, Seeds of Hunteria *(Abere)*, Bitter leaves *(Ewuro)*, Moringa *(Ewe-Ile)*, Scent Leaf *(Efinrin)*, Dutchman's Pipe *(Ako-Igun)*.
7	Diarrhea	Latex *(Abara)*, Connessi *(Irena)*, Tea bush *(Efinrin-Nla)*, Guava *(Gilofa)*, African rose Wood *(Gbingbin)*, Serpent wood *(Asofeyeje)*, Sterculia gum *(Ose-Awere)*, Garlic *(Ayuu)*, Bitter leaf *(Ewuro)*, *Lemon juice (Ọsàn –Lakuregbe).*	Latex *(Abara)*, Connessi *(Irena)*, Tea bush *(Efinrin-Nla)*, Guava *(Gilofa)*, African rose Wood *(Gbingbin)*, Serpent wood *(Asofeyeje)*, Sterculia gum *(Ose-Awere)*, Garlic *(Ayuu)*, Bitter leaf *(Ewuro)*.	Latex *(Abara)*, Connessi *(Irena)*, Tea bush *(Efinrin-Nla)*, Guava *(Gilofa)*, African Rose Wood *(Gbingbin)*, Serpent wood *(Asofeyeje)*, Garlic *(Ayuu)*.
8	Eczema	Prickly Amaranth *(Dagunro)*, Physic nut *(Lapalapa)*, Mistletoe (*Afomo-Oni Sana*).	Prickly Amaranth *(Dagunro)*, Physic nut *(Lapalapa)*.	Prickly Amaranth *(Dagunro)*, Physic nut *(Lapalapa)*.
9	Fibroid	Wonderful kola *(Obi-Ata)*, Mistletoe (*Afomo-Oni Sana*), Masquerade stick	Wonderful kola *(Obi-Ata)*, Mistletoe (*Afomo-Oni Sana*).	Wonderful kola *(Obi-Ata)*, Mistletoe (*Afomo-Oni Sana*).

S/No.	Disease/ Ailment	Medicinal Plants for Treatment after the First Round of Interview	Medicinal Plants for Treatment after the Second Round of Interview	Medicinal Plants for Treatment after the Third Round of Interview
		(Atori).		
10	Gonorrhea	Indian hemp *(Igbo)*, Baboon vine *(Shaworo)*, Masquerade stick *(Atori)*, African Parquetina *(Ogbo)*, Bowstring hemp *(Oja-Ikoko)*, Sphenocentrum *(Akerejupon),* Lime *(Osan-Wewe)* Wonderful kola *(Obi-Ata)*, Mistletoe (*Afomo-Oni Sana*).	Indian hemp *(Igbo)*, Baboon vine *(Shaworo)*, Masquerade stick *(Atori)*, African Parquetina *(Ogbo)*, Bowstring hemp *(Oja-Ikoko)*, Sphenocentrum *(Akerejupon),* Lime *(Osan-Wewe)* Wonderful kola *(Obi-Ata).*	Indian hemp *(Igbo)*, Baboon vine *(Shaworo)*, Masquerade stick *(Atori)*, Ogbo *(African Parquetina)*, Bowstring hemp *(Oja-Ikoko)*, Sphenocentrum *(Akerejupon),* Lime *(Osan-Wewe).*
11	Hernia	Mahogany *(Apa-Igbo)*, Tassel flower *(Odundun-Owo)*, Fertility tree *(Akoko)*, Moringa tree *(Ewe-Ile)*, Masquerade stick *(Atori).*	Mahogany *(Apa-Igbo)*, Tassel flower *(Odundun-Owo)*, Fertility tree *(Akoko)*, Masquerade stick *(Atori).*	Mahogany *(Apa-Igbo)*, Tassel flower *(Odundun-Owo)*, Fertility tree *(Akoko).*
12	Hypertension	Moringa tree *(Ewe-Ile)*, Adenia *(Arokeke)*, Cotton *(Owu)*, Garlic *(Ayuu)* and Ginger *(Ata-Ile)*, Mistletoe (*Afomo-Oni Sana*).	Moringa tree *(Ewe-Ile)*, Adenia *(Arokeke)*, Cotton *(Owu)*, Garlic *(Ayuu)* and Ginger *(Ata-Ile).*	Moringa tree *(Ewe-Ile)*, Adenia *(Arokeke)*, Cotton *(Owu)*, Garlic *(Ayuu)* and Ginger *(Ata-Ile).*
13	Erectile Dysfunction	Climbing black pepper *(Iyere)*, Cabbage tree *(Shapo)*, Crimson thyme *(Amuje wewe)*, Corynanthe yohimbe *(Ako Idagbon)*, Bushveld *(Ajekobale)*, Bushveld *(Ajekobale)*, Bowstring Hemp *(Oja-Ikoko)*, Masquerade stick *(Atori)*, Wonderful kola *(Obi-Ata).*	Climbing Black Pepper *(Iyere)*, Cabbage Tree *(Shapo)*, Crimson Thyme *(Amuje Wewe)*, Corynanthe Yohimbe *(Ako Idagbon)*, Bushveld *(Ajekobale)*, Bushveld *(Ajekobale)*, Bowstring Hemp *(Oja-Ikoko)*, Wonderful Kola *(Obi-Ata).*	Climbing Black Pepper *(Iyere)*, Cabbage Tree *(Shapo)*, Crimson Thyme *(Amuje Wewe)*, Corynanthe Yohimbe *(Ako Idagbon)*, Bushveld *(Ajekobale)*, Bushveld *(Ajekobale)*, Bowstring Hemp *(Oja-Ikoko).*
14	Malaria	Baobab *(Igi-Ose)*, Alligator pepper *(Ata-Ire)*, Yellow Allamanda *(Ododo Alamanda)*, Tallow tree *(Orogbo Erin)*, Lemon Grass *(Kooko Oba)*, Cashew *(Kasu)*, Hunteria umbellate *(Abeere)*, Mango *(Mangoro)*, Stool wood *(Ahun)*, Brimstone tree *(Oruwo)*, Tumeric *(Laali-pupa)*, Dragon's Blood Tree *(Elepo, Asunje, Otori)*, Serpent Wood *(Asofeyeje)*, Guava tree *(Gilofa*), African Yellow Wood *(Osopupa).*	Baobab *(Igi-Ose)*, Alligator pepper *(Ata-Ire)*, Yellow Allamanda *(Ododo Alamanda)*, Tallow tree *(Orogbo Erin)*, Lemon Grass *(Kooko Oba)*, Cashew *(Kasu)*, Hunteria umbellate *(Abeere)*, Mango *(Mangoro)*, Stool wood *(Ahun)*, Brimstone tree *(Oruwo)*, Dragon's Blood Tree *(Elepo, Asunje, Otori)*, Serpent Wood *(Asofeyeje)*, Guava tree *(Gilofa*), African Yellow Wood *(Osopupa).*	Baobab *(Igi-Ose)*, Yellow Allamanda *(Ododo Alamanda)*, Tallow tree *(Orogbo Erin)*, Lemon Grass *(Kooko Oba)*, Cashew *(Kasu)*, Hunteria umbellate *(Abeere)*, Mango *(Mangoro)*, Stool wood *(Ahun)*, Brimstone tree *(Oruwo)*, Dragon's Blood Tree *(Elepo, Asunje, Otori)*, Serpent Wood *(Asofeyeje)*, Guava tree *(Gilofa*), African Yellow Wood *(Osopupa).*
15	Pile	Crimson thyme *(Amuje Wewe)*, Balloon vine *(Shaworo)*, Bushveld *(Ajekobale)*, Asthma Plant *(Oroigi)*, *Lemon juice (Ọsàn –Lakuregbe)*, Bitter leaf *(Ewuro).*	Crimson Thyme *(Amuje Wewe)*, Balloon Vine *(Shaworo)*, Bushveld *(Ajekobale)*, Asthma Plant *(Oroigi)*, Bitter Leaf *(Ewuro).*	Crimson Thyme *(Amuje Wewe)*, Balloon Vine *(Shaworo)*, Bushveld *(Ajekobale)* and Asthma Plant *(Oroigi).*
16	Rheumatism	Crab's Eye *(Oju Ologbo)*, Acalypha (*Jinwinini*), Starburr *(Dagunro-gogoro)*, Alligator pepper *(Ata-Ire)*, Balloon vine *(Shaworo)*, Christmas Bush *(Ipa)*, Frest crinum (*Ogede-Odo*), African Yellow Wood *(Osopupa)*, Bushveld *(Ajekobale)*	Crab's Eye *(Oju Ologbo)*, Acalypha (*Jinwinini*), Starburr *(Dagunro-gogoro)*, Alligator pepper *(Ata-Ire)*, Balloon vine *(Shaworo)*, Christmas Bush *(Ipa)*, Frest crinum (*Ogede-Odo*), Bushveld *(Ajekobale)*	Crab's Eye *(Oju Ologbo)*, Acalypha (*Jinwinini*), Starburr *(Dagunro-gogoro)*, Alligator pepper *(Ata-Ire)*, Balloon vine *(Shaworo)*, Christmas Bush *(Ipa)*, Frest crinum (*Ogede-Odo*).
17	Ring Worm	Prickly Amaranth *(Dagunro)*, Physic nut *(Lapalapa)*, Masquerade stick *(Atori).*	Prickly Amaranth *(Dagunro)*, Physic nut *(Lapalapa).*	Prickly Amaranth *(Dagunro)*, Physic nut *(Lapalapa).*
18	Stomach Disorder	Rough-Chaff herb *(Aboro)*, Mahogany *(Apa-Igbo)*, Hemorrhage plant *(Yunyun)*, Corpse awakener *(Tude)*, Bush candle tree *(Origbo)*, Sandpaper tree *(Epin)*, Starburr *(Dagunro-Gogoro)*, *Lemon juice* (*Ọsàn –Lakuregbe*), Masquerade stick *(Atori).*	Rough-Chaff herb *(Aboro)*, Mahogany *(Apa-Igbo)*, Hemorrhage plant *(Yunyun)*, Corpse Awakener *(Tude)*, Bush Candle Tree *(Origbo)*, Sand Paper Tree *(Epin)*, Starburr *(Dagunro-Gogoro)*, *Lemon juice* (*Ọsàn –Lakuregbe).*	Rough-Chaff herb *(Aboro)*, Mahogany *(Apa-Igbo)*, Hemorrhage plant *(Yunyun)*, Corpse Awakener *(Tude)*, Bush Candle Tree *(Origbo)*, Sand Paper Tree *(Epin)*, Starburr *(Dagunro-Gogoro).*
19	Syphilis	Acalypha *(Jinwinini)*, Bear's breech *(Ahon-Ekun)*, Akerbia *(Dodo)*, Neem tree *(Eke-Oyibo)*, Pawpaw *(Ibepe)*, Silk cotton tree *(Araba)*, Starthistle *(Danyi)*, Dutchman's Pipe *(Ako-Igun)*, Garlic *(Ayuu)*, Masquerade stick *(Atori)*, Wonderful kola *(Obi-Ata).*	Acalypha *(Jinwinini)*, Bear's breech *(Ahon-Ekun)*, Akebia *(Dodo)*, Neem tree *(Eke-Oyibo)*, Pawpaw *(Ibepe)*, Silk cotton tree *(Araba)*, Starthistle *(Danyi)*, Dutchman's Pipe *(Ako-Igun)*, Garlic *(Ayuu)*, Wonderful kola *(Obi-Ata).*	Acalypha *(Jinwinini)*, Bear's breech *(Ahon-Ekun)*, Akebia *(Dodo)*, Neem tree *(Eke-Oyibo)*, Pawpaw *(Ibepe)*, Silk cotton tree *(Araba)*, Starthistle *(Danyi)*, Dutchman's Pipe *(Ako-Igun)*, Garlic *(Ayuu).*

S/No.	Disease/ Ailment	Medicinal Plants for Treatment after the First Round of Interview	Medicinal Plants for Treatment after the Second Round of Interview	Medicinal Plants for Treatment after the Third Round of Interview
20	Typhoid	Cashew *(Kasu)*, Pineapple *(Ope-Oyinbo)*, Dutchman's pipe *(Ako-igun)*, Bitter orange *(Orombo-Igun)*, African yellow wood *(Osopupa)*, Lecaniodiscus *(Akika)*, Scent Leaf *(Efinrin-Nla)*, Lemongrass *(Kooko Oba)*, Neem tree *(Dongoyaro)*, Ginger (*Ata-Ile)*, Pawpaw *(Ibepe)*, Bitter leaf *(Ewuro)*, Balloon vine *(Shaworo)*.	Cashew *(Kasu)*, Pineapple *(Ope-Oyinbo)*, Dutchman's pipe *(Ako-igun)*, Bitter orange *(Orombo-Igun)*, African yellow wood *(Osopupa)*, Lecaniodiscus *(Akika)*, Scent Leaf *(Efinrin-Nla)*, Lemongrass *(Kooko Oba)*, Neem tree *(Dongoyaro)*, Ginger (*Ata-Ile)*, Pawpaw *(Ibepe)*, Bitter leaf *(Ewuro)*.	Cashew *(Kasu)*, Pineapple *(Ope-Oyinbo)*, Dutchman's pipe *(Ako-igun)*, Bitter orange *(Orombo-igun)*, African yellow wood *(Osopupa)*, Lecaniodiscus *(Akika)*, Scent Leaf *(Efinrin-Nla)*, Lemongrass *(Kooko Oba)*, Neem tree *(Dongoyaro)*, Ginger (*Ata-Ile)*, Pawpaw *(Ibepe)*.
21	Ulcer	Crab's eye *(Oju-Ologbo)*, Acalypha *(Jinwinini)*, Bhadram *(Ewe-Owo)*, Goat Weed *(Imi-Esu)*, Prickly Amaranthus *(Tete Elegun-un)*, Dutchman's pipe *(Ako-Igun)*, Starburr *(Dangunro-Gogoro)*.	Crab's eye *(Oju-Ologbo)*, Acalypha *(Jinwinini)*, Bhadram *(Ewe-Owo)*, Goat Weed *(Imi-Esu)*, Prickly Amaranthus *(Tete Elegun-un)*, Dutchman's pipe *(Ako-Igun)*.	Crab's eye *(Oju-Ologbo)*, Acalypha *(Jinwinini)*, Bhadram *(Ewe-Owo)*, Goat Weed *(Imi-Esu)*, Prickly Amaranthus *(Tete Elegun-un)*.
22	Yellow Fever	Starburr *(Dangunro-Gogoro)*, Stool wood *(Ahun)*, African Rosewood *(Abara)*, Day Flower *(Itopere)*, Turmeric *(Laali-pupa)*, Fertility tree *(Akoko)*, Clausena *(Oboku)*, Morning glory plant *(Ododo-odo)*, Bitter kola *(Orogbo)*, Tallow tree *(Orogbo Erin)*, Garlic *(Ayuu)*, Silk cotton tree *(Araba)*.	Starburr *(Dangunro-Gogoro)*, Stool wood *(Ahun)*, African Rosewood *(Abara)*, Day Flower *(Itopere)*, Turmeric *(Laali-pupa)*, Fertility tree *(Akoko)*, Clausena *(Oboku)*, Morning glory plant *(Ododo-odo)*, Bitter kola *(Orogbo)*, Tallow tree *(Orogbo Erin)*, Garlic *(Ayuu)*.	Starburr *(Dangunro-Gogoro)*, Stool wood *(Ahun)*, African Rosewood *(Abara)*, Day Flower *(Itopere)*, Turmeric *(Laali-pupa)*, Fertility tree *(Akoko)*, Clausena *(Oboku)*, Morning glory plant *(Ododo-odo)*, Bitter kola *(Orogbo)*, Tallow tree *(Orogbo Erin)*.

5.0 CONCLUSION

The creation of a consensual knowledge base of medicinal plants and herbs for the treatment and/or management of various diseases in the ATM practice using the Delphi method yielded robust, holistic and non-subjective results. This is because it provided means for the harmonization of the ideas of the various experts recruited for the study. This will have been otherwise impossible if a single interview was conducted for all experts or multiple interviews conducted for a single expert. This was made possible through the elimination of medicinal plants and herbs which are not unanimously agreed by the experts as useful in the treatment and management of specific diseases and ailments. The elicited knowledge was documented as given in Table 1.5 for the use of both experts and non-experts in the field of study.

In conclusion, the application of the Delphi technique in the creation of a knowledge base of medicinal plants and herbs in the ATM practice achieved a consensual, holistic and non-subjective knowledge base of medicinal plants and herbs. The results obtained from this study will aid the development of an online ATM intervention designed to support the diagnosis and treatment of common ailments, summarize expert-driven consensus of medicinal plants and herbs, and direct future researches in the field of ATM practice.

REFERENCES

1. Beck, H., Dao-Tran, M., and Eiter, T. (2018). LARS: A Logic-based Framework for Analytic Reasoning over Streams. *Artificial Intelligence*, 261, pp. 16-70.
2. *Sharples*, S. *Shadbolt*, N., and *Smart*, P. R. (*2015*) *Knowledge Elicitation:* Methods, Tools, and Techniques. *In, Wilson*, John *R* and *Sharples*, Sarah (*eds.*) *Evaluation* of *Human Work. Boca Raton, Florida, USA. CRC Press,* pp. 163-200.
3. Gavrilova, T. and Andreeva, T. (2012). Knowledge Elicitation Techniques in a Knowledge Management Context. *Journal of Knowledge Management*, 16(4), pp. 523-537, https://doi.org/10.1108/13673271211246112.
4. Wellsandt, S., Hribernik, K. and Thoben, K. (2014). Qualitative Comparison of Requirements Elicitation Techniques that are used to collect Feedback Information about Product Use. In: *Proceedings of 24th CIRP Design Conference*, pp. 212–217.
5. Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M and Marinescu, V. (2013). Converting Unstructured and Semi-Structured Data into Knowledge. *Proceedings - RoEduNet IEEE International Conference*, pp. 1-4. DOI: 10.1109/RoEduNet.2013.6511736.
6. Scheurwegs, E., Luyckx, K., Luyten, L., Daelemans, W., and Van den Bulcke, T. (2016). Data Integration of Structured and Unstructured Sources for assigning Clinical Codes to Patient Stays. *Journal of the American Medical Informatics Association: JAMIA*. 23(1), pp. 11–19. DOI:10.1093/jamia/ocv115
7. Venkatavara P. D., Madhusudanan, S. and Jaganathan, S. (2015). uCLUST-A New algorithm for Clustering Unstructured Data. ARPN *Journal of Engineering and Applied Sciences*. 10, pp. 2108-2117.
8. Gharehchopogh, F. S. and Khalifehlou, Z. A. (2011). Analysis and evaluation of unstructured data: Text mining versus natural language processing. *5th International Conference on Application of Information and Communication Technologies 2011, AICT 2011*, pp. 1 - 4. DOI: 10.1109/ICAICT.2011.6111017.
9. Agbor, A. and Naidoo, S. (2016). A Review of the role of African Traditional Medicine in the Management of Oral Diseases. *African Journal of*

Traditional, Complementary and Alternative Medicines. 13, pp. 133 - 141. DOI: 10.4314/ajtcam. v13i2.16.

10. McMillan, S. S., King, M., and Tully, M. P. (2016). How to use the Nominal Group and Delphi Techniques. *International Journal of Clinical Pharmacy*, 38(3), pp. 655–662. DOI:10.1007/ s11096-016-0257-x.
11. Habibi A., Sarafrazi A., Izadyar S. (2014). Delphi Technique Theoretical Framework in Qualitative Research. *Int J Eng Sci*. 2014, 3(4), pp. 8-13.
12. Eddy, Matthew Daniel (2013). The Shape of Knowledge: Children and the Visual Culture of Literacy and Numeracy. *Science in Context*. 26(2), pp. 215–245. DOI:10.1017/s0269889713000045.
13. Audi R. (2010). Epistemology: A Contemporary Introduction to the Theory of Knowledge. 3rd Edition, Routledge, New York. ISBN-13: 978-0415879231
14. Girardi D., Kueng J., Holzinger A. (2015). A Domain-Expert Centered Process Model for Knowledge Discovery in Medical Research: Putting the Expert-in-the-Loop. In: Guo Y., Friston K., Aldo F., Hill S., Peng H. (eds) *Brain Informatics and Health. BIH 2015. Lecture Notes in Computer Science*, 9250. Springer, Cham. https://link. springer.com/chapter/10.1007/978-3-319-23344-4_38. Assessed on 17th October 2019.
15. Shadbolt, N. R., and Smart, P. R. (2015). Knowledge Elicitation. In J. R. Wilson & S. Sharples (Eds.), *Evaluation of Human Work (4th ed.)*. CRC Press, Boca Raton, Florida, USA. (http://www.amazon.co.uk/Evaluation-Human-Work-FourthWilson/dp/1466559616/).
16. O’Hagan, A. (2012). Probabilistic Uncertainty Specification: Overview, Elaboration Techniques and their Application to a Mechanistic Model of Carbon Flux. *Environmental Modelling and Software - ENVSOFT*. 36. DOI:10.1016/j. envsoft.2011.03.003.
17. Kleiter, G. D. (2018). Imprecise Uncertain Reasoning: A Distributional Approach. *Frontiers in Psychology*, 9, 2051 -2070.
18. O’Hagan, A. (2019). Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, 73(sup1), pp. 69-81, DOI: 10.1080/00031305.2018. 1518265.
19. Aniba, M. R., Siguenza, S., Friedrich, A., Plewniak, F., Poch, O., Marchler-Bauer, A., and Thompson, J. D. (2009). Knowledge-Based Expert Systems and a Proof-of-Concept Case Study for Multiple Sequence Alignment Construction and Analysis. *Briefings in Bioinformatics*, 10(1), pp. 11–23. DOI:10.1093/bib/ bbn045.
20. Papautsky, E. L., Crandall, B., Grome, A., and Greenberg, J. M. (2015). A Case Study of Source Triangulation: Using Artifacts as Knowledge Elicitation Tools in Healthcare Space Design. *Journal of Cognitive Engineering and Decision Making*. 9(4), pp. 347–358. DOI: https://doi.org/10.1177/ 1555343415613720.
21. Raghav, P. R., Kumar, D., and Bhardwaj, P. (2016). Experience of Delphi Technique in the Process of Establishing Consensus on Core Competencies. *International Journal of Applied & Basic Medical Research*. 6(3), pp. 191–194. DOI:10.4103/2229-516X.186966
22. Bentley M., Kerr R. and Powell S. (2016). The Use of a Modified Delphi Technique to Inform the Development of Best Practice in Inter-Professional Training for Collaborative Primary Healthcare. *Journal of Research in Inter-Professional Practice and Education*. Vol. 6. Assessed on 10th September 2018. DOI: http://dx.doi.org/10.22230/jripe.2016 v6n1a 232.
23. Awoniran, O. M., Soriyan, H. A. and Elujoba, A. A. (2015). A Framework for Knowledge Capture in African Traditional Treatment of Malaria. *Nigerian Journal of Natural Products and Medicine*. Vol. 19, Pp 126-133. DOI: http://dx.doi.org/10.4314/njnpm .v19i1.15.
24. Akinyemi, O. V. and Olatokun W. M. (2015). Factors Influencing Indigenous Knowledge Data Elicitation from Herbal Medicinal Practitioners in South-Western Nigeria. *African Journal of Sustainable Development*. 5(1), pp 97-121. ISSN: 2315-6317.
25. Elujoba, A. A., Odeleye, O. M. and Ogunyemi C. M. (2005). Traditional Medicine Development for Medical and Dental Primary Health Care Delivery System in Africa. *African Journal of Traditional, Complementary and Alternative Medicines*, 2(1), pp. 46-61 ISSN: 0189-6016.
26. World Health Organization (2000). *General Guidelines for Methodologies on Research and Evaluation of Traditional Medicine*. WHO/EDM/ TRM/2000.1, pp. 1-74.
27. Paulo P. M., Kofi B. and Ossy M. K. (2010). Clinical Practices of African Traditional Medicine. *African Health Monitor*, Issue 13, World Health Organization Africa. https://www.aho.afro.who. int/en/ahm/issue/13/reports/clinical-practices-african-traditional-medicine. Accessed on 10th May 2017.
28. Gwet, K. L. (2008). Computing Inter-Rater Reliability and its Variance in the Presence of High Agreement. *British Journal of Mathematical and Statistical Psychology*, 61, pp 29–48.
29. Marozzi, Marco (2014). Testing for Concordance between Several Criteria. *Journal of Statistical Computation and Simulation*. 84(9), pp. 1843–1850. DOI:10.1080/00949655.2013.766189.
30. Corder, G.W., Foreman, D.I. (2009). Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach. *Wiley*, ISBN 978-0-470-45461-9.

UTILIZATION OF ICT FACILITIES IN TEACHING BIOLOGY IN SENIOR SECONDARY SCHOOLS IN GOMBE STATE

[1]Yilshik O. M., [2]Ezekiel D. P. and [3]Daniel H.
Department of Science Education, Faculty of Education, Federal University of Kashere, Gombe State
[1]oyilshik@gmail.com, +2348032904643; [2]daziezekiel@gmail.com, +2348037668529; [3]yankuril@yahoo.com, +234802 799 3733

ABSTRACT

This study examined the utilization of Information and Communication Technology in senior secondary school biology teachers in Gombe state. Descriptive survey research design was employed. Three research questions were raised and one hypothesis was tested at 0.05 alpha level. The population comprised all the 225 Biology Teachers in both public and private Senior Secondary Schools in Gombe state. 12 public and 8 private Secondary schools were purposively sampled from each of the three senatorial zones of the state. 2 Biology Teachers were randomly sampled from each school and made a sample of 120 respondents from 60 senior secondary schools. Biology Teachers' Utilization of ICT Facilities Questionnaire (BTUICTFQ) was used as instrument for data collection. The data collected were analysed using frequency tables, simple percentages, mean, standard deviation and t-test. Results showed that the common ICT facilities available includes; computers, mobile phone, printers, and computer laboratories. The findings also revealed that the available facilities were underutilized. The hypothesis tested revealed that there is no significant difference in the mean score of teachers' response on the utilization of ICT facilities in public and private senior secondary schools. Hence, the study recommended, training and re-training of biology teachers on ICT, create curriculum policies to integration ICT in teaching and learning process and provide incentives of ICT gadgets.

Keywords: ICT facilities, Utilization, Biology, Teaching and Learning.

1.0 INTRODUCTION

Scientific knowledge is rapidly expanding with the use of Information and Communication Technology (ICT) resources and services (Etiubun & Akpan, 2017). These resources include digital cameras, computers, multimedia software applications, internet, television video. Biology is one of the subjects offered by senior secondary school students in Gombe state, Nigeria. It is one of the core subjects of the curriculum as reflected in the 6th edition of the National Policy on Education (FRN, 2013).

Roger, Joshi and Green, (2019) defined biology as the study of living things and their vital processes. The field deals with all the physiochemical aspects of life. Biology is both fascinating and fun to study at the elementary and secondary levels, especially when connected to everyday life. The subject biology piques intellectual curiosity, increases awareness of fragile ecosystem and stimulates critical thinking (Martens, 2018). Some of the aims and objective of teaching biology in senior secondary schools aimed at increasing understanding of living systems and allows the opportunity to compare the systems in relationship to the self and other organisms in the natural environment. This implies that studying biology is essential for job preparation such as working in either an educational institution or industry concerned with research and development of drugs, food-related items and biotechnology. Biology also trains individuals in skills needed to be functionally relevant in solving practical societal and global problems which includes: combating desert encroachment, global warming, environmental pollution or degradation, protection of endangered species, maintaining a balanced ecosystem and so on.

Information and communication Technology (ICT) refers to technologies that provide access to information through telecommunications (Christenson, 2010). This indicates that ICT is a range of technology for gathering, storing, retrieving, processing, analysing, and transmitting information. It focuses primarily on communication technologies which includes the internet, wireless networks, cell phones, and other communication mediums. Modern communication technologies have created a global village, in which people can communicate with others across the world in a very easy way. For this reason, ICT is often studied in the context of how modern communication technologies affect society.

In the words of Apagu and Wakili (2015), the importance of ICT in the teaching and learning cannot be overemphasised. This is because its adoption by teachers will enhance effective teaching. A lot of studies revealed the relevance of ICT in science teaching and learning. Bature, (2016) found that effective use of ICT tools in teaching and learning of mathematics improves students' performance, enhances teaching and learning, and also makes students interested in learning. In the same vein, Anderson, (2018) posited that despite the abstract nature of physics, through the use of ICT, the teaching and learning of physics has been made interesting by making the otherwise abstract matter concrete. It seems that the available ICT facilities are underutilised for teaching biology. It therefore, pertinent to investigate the utilization of ICT in the teaching and learning of biology in senior secondary schools in Gombe State.

The objectives of this research are therefore, to find out:

(a) The ICT facilities available in senior secondary schools in Gombe state

(b) Whether biology teachers utilize ICT facilities in teaching biology in senior secondary schools in Gombe state?
(c) The barriers to effective utilization of ICT facilities by biology teachers in senior secondary schools in Gombe state.
(d) Whether any difference exist between teachers' utilization of ICT facilities in public and private senior secondary schools in Gombe state.

2.0 LITERATURE REVIEW

Information and communication Technology (ICT) is defined by Vijayalakshmi, (2018) as a divers set of technological tools and resources used to communicate, to also create, store, disseminate, and manage information. Vijayalakshmi maintained that ICT comprised all the technology used to handle telecommunications, broadcast media, Intelligent building management systems, audio-visual processing and transmission systems and network-based control and monitoring functions. It is therefore seen as one of the indispensable tools or powerful force of achieving effective teaching and learning (classroom delivery) in our educational system. Ikemelu (2015) sees ICT as a new approach for science teaching and learning which is necessary to create more access to information and experience through global networks and pool of knowledge. He further asserted that a good science teacher (and Biology teacher in particular) who must be functionally relevant in the 21st century is expected to acquire ICT skills and competencies in basic ICT computer applications software such as MS-word, CorelDraw, MS-excel, MS-access, operating system E-mail, power point and so on. It promotes learning, motivates and empowers the learning as well as facilitates the job of the teacher (UNDP, 2006). ICT has the potentials to improve teaching and enhance institutional administration (Kazu & Yavulzalp, 2008).

Availability of ICTs in schools however will not guarantee their effective use. The key to how those tools are used is the teacher. The teacher is defined as a person who has acquired the requisite knowledge, pedagogical skills with appropriate level of education in Nigeria (Ciwar, 2006). He is seen as a facilitator of learning and a person who never ceases to be a learner, continue to promote the spirit of enquiry, cultivate creativity, innovation and divergent thinking. A biology teacher trains an individual in critical thinking and application of scientific principles in solving practical problems for useful living and effective management of the ecosystem. The Federal Government of Nigeria (F. R. N, 2004), recognizes the prominent role of ICT in the modern world and has integrated ICT into education in Nigeria. But the uptake in school is still very low. Onuka, Ajayi, and Lawani (2013) observed that most schools both private and public secondary schools do not offer ICT training programmes. This implies that computer is not part of the classroom technology in more than 80% of Nigeria secondary schools, that means chalkboard and textbook continue to dominate classroom activities in most Nigeria secondary schools.

E-learning is basically the use of information and communication technologies (ICTs) to enhance teaching and learning and research (Eteng & Ntui, 2009). In this age of Information and Communication Technology (ICT), there is growing concern for the use of ICT resources such as the computer, scanner, printer, Intranet, Internet, e-mail, videophone systems, teleconferencing devices, wireless application protocols (WAP), radio and microwaves, television and satellites, multimedia computer and multimedia projector, among others in instructional delivery method (Nwana, 2012). Inije, Utoware, and Kren, (2013) posited that with e-learning, there is a shift from the traditional approach of teacher-directed didactic to modern methods where computer technology plays a significant role, thereby improving the quality, efficiency and effectiveness of teaching, learning, research and educational management. They further explained that E-learning as a sub-system within ICT, is the electronic process which enhances the delivery and administration of learning opportunities and support by using computer, networked and web-based technology to help individual performance and development. The basic principle of e-learning is connectivity which is the process by which computers are networked to share information that can connect people. This is provided for by what is often called the e-learning landscape or architecture, which refers to the hardware, software and connectivity components required to facilitate learning. Therefore, the use of information and technology in education is one way through which learner's achievement can be enhanced.

In line with changes in the society, the 21st century economy requires school graduates to be well conversant with modern electronic knowledge, embracing computer technology and other forms of media so as to remain competitive in the global employment market. This indicates that learners without literacy in computer technology are at risk of being left out of global employment opportunity. Learning institutions therefore have opportunity to improve performance skills and enter into the service sector economy by adopting use of computer technology. This will ensure that learners have the skills required for the ever increasingly competitive labour market (Gideon, Samwel, Ronnie, & John, 2017). Since students in Nigeria are part of a digital society with access to a wide range of information provided using information and communication technology (ICT), it is important to include ICT in the educational (Teaching and learning) process as well. In order to make this possible, schools should not only have the necessary infrastructure and adequate space, but also computer literate teachers who are able to create diverse teaching and learning opportunities through ICT in the classrooms (Bulic, & Novoselic, 2014). According to Ajayi (2008), the use of these facilities involves various methods which include systematic feedback system, computer-based operation network, video conferencing and audio conferencing, internet worldwide websites and computer assisted instruction. This delivery method increases the possibilities for how, where and when learners can engage in lifelong learning. This is a clear indication that E- learning as a sub-set of ICT has the potential to transform how and when learners learn.

Utilization simply put, refers to practical and effective use of something. The call for utilization of ICT in teaching Biology is to infuse and inject efficiency and effectiveness in curriculum implementation. Over the years, most public and private secondary schools seem to have neglected the utilization of ICT facilities for teaching. Asaolu and Fashanu (2012), established in their studies that while there is a remarkable improvement in ICT status of public schools over the decade, they still lag behind considerably when compared to private schools. However, in a developing country as Nigeria, e-learning is challenged with the problem of material devices such as computer, computer laboratories, internet and e-mail facilities, videophone systems and teleconferencing devices, fax and wireless applications, digital library, digital classrooms, multimedia systems and the problem of multimedia and courseware development. Other studies indicated that there is dearth of trained teachers for e-learning, lack of facilities, infrastructures and equipment Jegede and Owolabi (in Inije, Utoware, & Kren, 2013). Albion, Tondeur, Forkosh-Baruch, and Peeraer, (2015) revealed in their studies that inadequate infrastructure and lack of training is a major impediment to integration of ICT in instructional processes. In the same vein, Agmmuoh (2015) identified some of the factors impeding the application of ICT in Nigerian secondary schools to include; lack of training among teachers, irregular and inadequate power supply, high cost of computer equipment and others like software, printers, scanners.

Empirical Review.

Ochu, (2016) investigated the availability and usability of ICT resources in classroom delivery by secondary school chemistry teachers in Benue state. Survey research design was adopted. 36 chemistry teachers were randomly sampled from all the chemistry teachers in the state and utilised as the sample for the study. A structured questionnaire was used to gather data. Percentages, means and standard deviations were applied to analyse the data. The research findings revealed that almost all the ICT facilities and resources for teaching Chemistry were not available in secondary schools used for the study. It was concluded that ICT resources were not being used in teaching chemistry in senior secondary schools in Benue state. A similar study was carried out by Livingstone (2015) on the use of ICT in the teaching and learning of physics in secondary schools in Kimili district of Bugoma county, Kenya. A descriptive survey design was adopted. 30 physics teachers and 250 form three students were sampled from 11 out of 23 secondary schools in the district and used as sample for the study. The study employed questionnaire, interview guide, lesson observation schedule and document analysis guide in data collection. Frequency counts, percentages, means and correlations were utilised in the analysis of the data collected and reported that Physics teachers mainly used ICT facilities at the planning stage of lessons, especially typing of lesson plans and scheme of work, computing and analyzing student test scores. However, ICT use during lesson delivery was minimal.

In addition, Bature, (2016) carried out a research on the role of ICT as a tool for effective teaching of Mathematics in Secondary Schools in Kafur Local Government Area of Katsina State. The study adopted survey research design. The population was the entire students and mathematics teachers in Kafur Local Government Area. Five out of the ten schools in the area of study were randomly selected as the study sample. Questionnaire was used to collect data from both students and teachers. Simple percentage and chi-square were used to analyse data. Among the findings are the effective use of ICT tools in teaching and learning of mathematics improve students' performance, enhance teaching and learning, and also make students interested in learning.

2.1 Statement of the Problem

Nigerian secondary educational system in recent years has witnessed berets of unimpressive performances in national and international examinations such as West Africa Examination Council (WAEC) and National Examination Council (NECO) most especially in science subjects including Biology. This position further buttressed by WAEC Senior Secondary Certificate Examination (SSCE) Chief Examiner's Report on biology for 2015, 2016 and 2018 showed that the overall performance of students in biology is poor. The report further revealed that school candidates' weaknesses that resulted to poor performance include poor drawing of diagrams, poor attempt in answering questions that require detailed explanations, wrong spelling of biological names and terms, inability to use technical terms to describe some processes, in ability to compare biological processes appropriately. The consistent poor performance of secondary school students in biology at the SSCE leaves one in doubt about the effectiveness of teaching methods and strategies popularly used by biology teachers for teaching the subject. The poor performance could be attributed to poor and inappropriate instructional methods and under-utilization of instructional materials including ICT facilities.

Although Gombe state government, Parent-Teacher Association, and Old Students Associations have made appreciable efforts to stem the tide by building and equipping computer laboratories in a number of the secondary schools across the state, it has not yielded the needed result. It seems therefore that the poor performance could be due to underutilization of ICT facilities in teaching. It becomes pertinent to find out whether or not, biology teachers use ICT facilities in teaching and learning biology in senior secondary schools Gombe State.

2.2 Research Questions

The research sought to answer the following questions:

(a) What are the ICT facilities available in senior secondary schools in Gombe state?

(b) Do biology teachers utilize ICT facilities in teaching and learning in senior secondary schools in Gombe state?

(c) What are the barriers to effective utilization of ICT facilities in teaching biology in senior secondary schools in Gombe state?

2.3 Research Hypothesis

This null hypothesis was formulated for testing at 0.05 level of significance:

HO: There is no significant difference in the mean score of teachers' response on the utilization of ICT facilities in public and private senior secondary schools.

3.0 METHODOLOGY

The study adopted a descriptive survey research design using a population of 225 biology teachers in both public and private secondary schools located in the three senatorial zones of the state. 12 public and 8 private secondary schools were purposively sampled from each senatorial zone. 2 biology teachers were randomly sampled from each school and made a sample of 120 respondents from 60 senior secondary schools.

3.1 Instrumentation

One researcher-constructed instrument referred to as Biology Teachers' Utilization of ICT Facilities Questionnaire (BTUICTFQ) was used for data collection. The BTUICTFQ was divided into four sections. Section A requested for the background information of the biology teachers, section B had 7-items on availability of ICT facilities with options of available, not available and available but in bad shape. and section C had 7-items on the utilization of ICT facilities and it is a four-point Likert scale with responses of strongly agree (4), Agree (3) Disagree (2) and strongly disagree (1), while section D had 7 -items on the barriers to effective utilization of ICT facilities in teaching and learning biology in senior secondary schools and it is also a four-point Likert scale with responses of strongly agree (4), Agree (3) Disagree (2) and strongly disagree (1). The content and construct validity was carried out by experts. The instrument was pilot tested outside the study sample and a reliability coefficient of 0.82 was obtained using Cronbach's Alpha.

4.0 DATA ANALYSIS

The data were analysed using descriptive and t-test statistics.

Research Question 1: What are the ICT facilities available in secondary schools in Gombe State?

Table 1: **ICT Facilities Available in Senior Secondary Schools**

S/No	ICT Equipment	Available (%)	Not Available (%)	Available but in Bad shape (%)
1.	Computer laboratory	50	30	20
2.	Computers	60	10	30
3.	Smartphones/mobile phones	70	5	25
4.	Internet connectivity	40	50	10
5.	DVD players	20	70	10
6.	Projectors	15	80	5
7.	Printers	60	15	25

Table 1 showed that ICT facilities such as smart-phones/mobile phones (70%), computers (60%), printer (60%) and computer laboratory (50%) are most available while internet (40%), DVD players (20%) and projectors (15%) are barely available. This implies that the ICT facilities readily available in secondary schools are; mobile phones, computers, printer and computer laboratories while projectors, DVD players and internet connectivity are not available.

Research Question 2: Do biology teachers utilize ICT facilities in teaching and learning in senior secondary schools in Gombe state?

Table 2: Utilization of ICT Facilities by Biology Teachers in Senior Secondary Schools

S/No	Items	Mean	SD	Remark
1	I use computer to teach biology when the need arise	2.53	.85	Agreed
2	I use projector for audio-visual transmission during my lesson	1.93	.48	Disagreed
3	I give my students on-line based assignments using Email	1.67	.51	Disagreed
4	I print out on-line resources for my students	3.08	.28	Agreed
5	I play biological documentaries for my students using DVD player	2.38	.78	Disagreed
6	I create WhatsApp group to share learning materials with my students.	2.13	.83	Disagreed
7	I source learning materials from the internet to teach my students.	3.30	.59	Agreed

Note: Criterion Mean= 2.50

Table 2 showed that the biology teachers agreed that they use computers to teach when the need arise, print out on-line resources for their students and source materials from the internet to teach their students as revealed in items 1, 4 and 7 with the mean scores of 2.53, 3.08 and 3.30 respectively. The result also clearly shows that the teachers disagreed that they use projectors, give students on-line assignments using E-mail, use DVD player to play biological documentaries and create WhatsApp to learning materials with students as shown in items 2, 3, 5, and 6 with corresponding mean scores of 1.93, 1.67, 2.38 and 2.13. This implies that most of the ICT facilities were not utilized in teaching biology in senior secondary schools in Gombe state.

Research Question 3: What are the barriers to effective utilization of ICT facilities in teaching biology in senior secondary schools in Gombe state?

The result in Table 3 shows that biology teachers agreed that inadequate computers in schools, lack of computer instructional software, irregular power supply and inadequate internet connectivity in schools are the major barriers to effective utilization of ICT facilities in teaching biology in Gombe state. Items 1, 2, 3, and 7 with corresponding mean scores of 2.57, 3.25, 3.55, 3.42 support that. The respondents also disagreed on teachers' lack of requisite skills, low interest in using ICT in teaching biology and school heads' lack of interest in ICTs as barriers to effective utilization of ICT facilities in teaching biology as seen in items 4, 5 and 6 with mean scores of 1.82, 1.88 and 1.68.

Table 3: Barriers to effective utilization of ICTs facilities in teaching biology

S/No	Items	Mean	SD	Remark
1	Inadequate computers in schools	2.57	.91	Agreed
2	Lack of computer instructional software	3.25	.80	Agreed
3	Irregular power supply to use ICT facilities	3.55	.49	Agreed
4	Teachers lack requisite knowledge and skills to apply ICTs in teaching biology	1.82	.47	Disagreed
5	Teachers' have low interest in using ICT facilities for teaching	1.88	.41	Disagreed
6	School heads are not interested in ICTs	1.68	.50	Disagreed
7	Inadequate internet connectivity in the schools	3.42	.49	Agreed

Note: Criterion Mean= 2.50

4.1 Hypothesis

HO: There is no significant difference in the mean score of teachers' response on the utilization of ICT facilities in public and private senior secondary schools.

Table 4: t-test of Teachers' Response on Utilization of ICT Facilities in Teaching Biology

School type	N	Mean	Std. Deviation	Std. Error Mean	Df	t	Sig.(2-tailed)
Public	60	4.27	1.247	.161	118	-1.386	.168
Private	60	4.63	1.626		.210		

From table 4, the calculated t = -1.386 < t- critical = 2.01 at .05 alpha level revealed that there is no significant difference in the mean score of teachers' response on the utilization of ICT facilities in public and private senior secondary schools. Hence, the null hypothesis is retained. This implies that teachers in both public and private secondary schools do not differ in their use of ICT facilities in teaching biology in Gombe state.

4.2 Discussion

Table 1 showed that ICT facilities readily available in the schools are; mobile phones, computers, printer and computer laboratories, while projectors, DVD players and internet c onnectivity are not available. This corroborates the findings of Sulaiman, Hindatu, and Lawal (2017) who found computers, television and CD player as the available ICT facilities in senior schools for teaching and learning biology. But this does not agree with Ochu, (2016) who revealed in his studies that most of the ICT facilities for chemistry teaching and learning were not available in secondary school. Table 2 revealed that most of the ICT facilities available in the sampled schools were not adequately utilized by biology teachers for teaching. The teachers only print out on-line resources, source materials from the internet and use computers to teach biology when the need arise not as an integral part of their teaching process. While projectors, E-mail, DVD player, WhatsApp groups were not utilized for teaching students This agrees with Sulaiman, Hindatu and Lawal, (2017) who found over 60% of biology teachers not using ICT for teaching. This has also buttressed the findings of Livingstone (2015) who revealed that Physics teachers mainly used ICT facilities at the planning stage of lessons, especially typing of lesson plans and scheme of work. They also used computers to analyse student test scores. However, ICT used during lesson delivery was minimal. Table 3 indicated that inadequate computers in schools, lack of computer instructional software, irregular power supply and inadequate internet connectivity in schools are the major barriers to effective utilization of ICT facilities in teaching biology secondary schools in Gombe state. This corroborated the findings of (Albion et al, 2015). They revealed that inadequate infrastructure and lack of training are major impediment to integration of ICT in instructional processes. In line this finding, Agommuoh (2015) lent credence and identified irregular and inadequate power supply, the high cost of computer software, printers, scanners as some of the factors impeding the use of ICT in education.

Table 4 showed that the null hypothesis was retained, implying that teachers in both public and private secondary schools do not differ significantly in their use of ICT facilities in teaching biology in Gombe state. Though the teachers in private secondary schools had a slightly higher mean score than their counterparts in public secondary schools, the difference was not significant. This does not agree with the finding of Asaolu, and Fashanu, (2012) their study established that while there was a remarkable improvement in ICT status of public schools over the decades, they still lag behind considerably when compared with private schools. It also contradicts the finding of Malero, Ismail, and Manyilizu, (2015). whose study revealed that private secondary schools in Dodoma municipality are better off than public secondary schools in ICT usage.

5.0 CONCLUSION AND RECOMMENDATIONS

The study showed a paucity of ICT facilities for teaching and learning biology in senior secondary schools in Gombe state. The few that were available were underutilized. In other words, most of the ICT facilities were mainly used for communication and administrative purposes. There is also limited connectivity, inadequate training of biology teachers, unstable source of power and insufficient software which are some of the barriers to effective utilization of ICT facilities. In both public and private senior secondary schools.

The following recommendations were made based on the findings of the study:

(a) More ICT facilities should be availed to secondary schools and located in science laboratories and staff rooms where the teachers can easily access them.

(b) Secondary school biology teachers should be given adequate training on ICT at the pre-service training institutions in order to acquire the knowledge and skills needed for effective utilisation of ICT facilities in the teaching and learning process.
(c) The Government should formulate clear policies to guide the process of ICT integration into the teaching and learning process in secondary schools.
(d) Government should sponsor secondary school biology teachers to develop science subject content materials and lesson notes for delivery through the internet, making and delivering CD-ROMs to the schools.
(e) Welfare packages for teachers such as laptops, tablets, modems and science allowance in other to encourage to utilized ICT facilities in teaching.
(f) Private and public secondary school biology teachers should be mandated for on-the-job ICT training and all facilities are to be made available for them.
(g) Every school should have an ICT coordinator.

REFERENCES

Agommuoh, P. C. (2015). Enhancing the teaching of physics through the use of ICT in senior secondary schools. *STAN 56th Annual Conference Proceedings*. 274-283. STAN. Abuja.

Ajayi, I.A. (2008). Towards effective use of information and communication technology for teaching in Nigerian colleges of education. *Asian Journal of Information Technology*, 7(5), 210 – 214.

Albion, P. R., Tondeur, J., Forkosh-Baruch, A. & Peeraer, J. (2015). Teachers' professional development for ICT integration: towards a reciprocal relationship between research and practice. *Education and Information Technologies*, 20 (4), 655-673.

Anderson, U, (2018). The Impact of Information and Communication Technology(ICT) on Teaching and Learning of Physics. http://afribary.com/works/the impact-of-information-and-communication-technology-ict-om-teaching-learning-of-physics-1256.

Apagu, V. V. & Wakili, B. A. (2015). Availability and utilization of ICT facilities for teaching and learning of vocational and technical education in Yobe state Technical Colleges. *American Journal of Engineering Research*, 4(2) 113-118. Retrieved from www.ajer.org

Bature, B. (2016). The role of information and communication technology as a tool for effective teaching and learning of mathematics. *Journal of Applied & Computational Mathematics.* 5: 333.doi: 10.4172/2168-9679.1000333

Buabeng-Andoh, C. (2012). An exploration of teachers' skills, perceptions and practices of ICT in teaching and learning in the Ghanaian second-cycle schools. *Contemporary Educational Technology*, 3 (1), 36-49.

Buabeng-Andoh, C. (2012). Factors influencing teachers' adoption and integration of information and communication technology into teaching: A review of the literature. *International Journal of Education and Development using Information and Communication Technology*, 8 (1), 136.

Bulic, M. & Novoselic, D. (2014). E-learning in a Biology Classroom. ICT in Practice. Retrieved from www.ictinpractice.com

Christennson, P. (2010). *ICT Definition*. Retrieved from https://techterms.com.

Ciwar, A. M. (2006). Teacher and competence and preparation for the 21st century. Retrieved from www.idea.info/document /paper 1162d3.pdf.

Eteng, U. & Ntui, I. A. (2009). Access to e-learning in the Nigerian University System (NUS): A case study of University of Calabar. The Information Technologies*: An International Journal of Information and Communication Technology*, 6(2) 1-10.

Etiubun, R. U., & Akpan, A. O. (2017). Science teachers' perception of ICT capacity building workshop in Askwa-Ibom state secondary schools, Nigeria. African Research Review: *International Multi-Disciplinary Journal, Ethiopia*,11(2).

Federal Republic of Nigeria. (2004). *National policy on education*. Lagos. NERDC press.

Federal Republic of Nigeria. (2013). *National policy on education*. Lagos. NERDC press.

Gideon M., Samwel M., Ronnie M., John M. (2017). Integrating ICT into Teaching and Learning Biology: A Case for Rachuonyo South Sub-County, Kenya. *American Journal of Education and Information Technologies*, 1(2),17-23. doi: 10.11648/j.ajeit.20170102.12

Ikemelu, C. R. (2015). Towards Effective Application of ICT Education for classroom curriculum delivery: Science Teacher Perspective. *STAN 56th Annual Conference Proceedings*. 330-338. STAN. Abuja.

Inije, G.O., Utoware, J. D. A. & Kren-ikidi, P.C. (2013). Utilization of E-learning Technologies in Business Education instructional delivery in colleges of Education in Delta state of Nigeria. *International Journal of Education and Research.* 1(10). Retrieved from www.ijern.com.

Kazu, I. Y., & Yavulzalp, N., (2008). An analysis of the primary school teachers' usage of instructional software. *International Journal of Emerging Technologies*,3 (1),45-53.

Livingstone, M.J. (2015). *Use of ICT in teaching physics*. A case of secondary school in Kimilili district, Bungoma county, Kenya. Unpublished master's thesis. Kenyatta University. Retrieved from https://www.semanticscholar.org.

Martens, H. (2018). *Aims and objectives of studying Biology*. Retrieved from https://www.theclassroom.com/aims-obj-study-biology-8449055.html.

Nwana, S. (2012). Challenges in the application of e-learning by secondary school teachers in Anambra State, Nigeria. *African Journal of Teacher Education*, 2(1), 67 -72.

Ochu, A.N.O (2016). Assessment of the availability and usability of information and communication technology (ICT) resources in classroom delivery by secondary school chemistry teachers. *Journal* *of Education Research and Behavioural Sciences*, 5(6), 094-097

Malero, A., Ismail, A. & Manyilizu, M. (2015). ICT usage readiness for private and public secondary schools in Tanzania, a case of Dodoma Municipality. *International Journal of Computer Application*, 129(3). Retrieved from www.ijcaonline.org .

Osaolu, O. S. & Fashanu, T. A., (2012). Adoption of ICT and its Comparative IMPACT on Private and Public schools in Lagos State, Nigeria. *International Journal of Emerging Technologies*, 3(1). Retrieved from http://excelingtech.co.uk/

Roger, K., Joshi, S. H, & Green, E. R., (2019). Encyclopaedia Britannica, Inc. Retrieved from https://www.britannica.org. science UNDP (2006). *Information communication technologies and development*. Retrieved from http://usdnhq.undp.prg/it4dev

Sulaiman, H., Hindatu, H. & Lawal, (2017). Teachers awareness on utilization of ICT facilities for Biology teaching in secondary schools in Matazu Local Government Area of Katsina State. *International Journal for Innovative Research in Multidisciplinary Field*, 3(10). Retrieved from www.IJIRMF.COM .

Vijayalakshmi, M, (2018). Information and communication technology in education. Retrieved from https://sweetmenthatab.tkcontact

West African Examination Council [WAEC], (2018). *Chief examiners' Reports*. Lagos. Retrieved from www.waeconline.org.ng/e-learning/Biology/Bio227mw.html.

RE-ENGINEERING OF A WEB-BASED LEAVE MANAGEMENT SYSTEM: DEVELOPMENT PERSPECTIVE

Ikono R. N., Olaiya B.O., Alade S. M., and [*]Gambo I. P.
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria

*Email of Corresponding Author: *ipgambo@yahoo.com*

ABSTRACT

The presence of a web-based leave management system has become a fundamental and ***i****ndispensable tool for every organization and business in this era of net-centric computing. It can manage and track multiple types of employee leaves and process request time away from work for employees. However, concerns about ineffectiveness, downtime, and delay in process and approval of leave request in the human resources department of organizations and agencies were identified. This paper seeks to explain the need to develop a system to remove the bottlenecks encountered in getting leave approved on time using web-technologies****.*** *This paper employed the use of a qualitative research approach to gain insights and familiarity with what exists to allow further study. Interviews, scenario-generation from stakeholders' and participants' experiential narratives and the analysis of current processes and structures were used as sources of data collection. A software development methodology was used in the development process. A web-based leave management system (WBLMS) was designed using Unified Modelling Language (UML) tools with its database designed using SQLite, implemented using web technologies and evaluated using DeLeon and McLean Information theory. The result obtained showed a positive outcome from the system evaluation. The paper concludes that a better leave system must be developed to enhance user satisfaction, increase productivity and ensure efficiency as well as the effectiveness of the process and employees while reducing the time taken in processing leave application****.***

Keywords: Delay, Efficiency, Effectiveness, Web-Technologies

1.0 INTRODUCTION

The concept of human resources is concerned about the process of dealing with people in the workplace. Human resource management (HRM) is the process that binds people with organization thereby helping the people and the organization to achieve their goals. HRM in organization and businesses in the world have become an indispensable arm of every organization to ever-increasing productivity of humans. This is because their functions are concerned with issues relating to compensation, performance, development, safety wellness, benefit, employee motivation and training (Griffith, 2018). With the advancement and improvement in Information and Communication Technology (ICT) in recent times, there is a paradigm shift in the consciousness of the role of Information and Communication Technology (IT) in human resource management. Today, several organizations and businesses have realized the growing importance of the use of IT in making optimum use of their HR functions. This comes in form of electronic human resource management, which depends on the state of–the–art technology extending from internet-based human resource management information system (HRIS) to shared and common intranet and enterprise portals (Mukherejee *et al.*, 2014; Rohilla, 2014).

Therefore, HRMIS can be described as a system that permits to keep track of your employees and information about them. In fact, Information technology (IT) and Human resource management have a myriad of effect on each other and therefore human resource personnel should be permitted to adopt the technological development that allows the re-engineering of the Human resource action, be prepared to maintain an organization and work project caused by the technology and be able to maintain an effective administrative atmosphere for innovation and knowledge-driven organization (Manpreet et al., 2014).

Within the field of ICT, the use of electronic or online WBLMS is to ensure that standard, effective management and productivity (physical and mental) of employees in the organization are maintained. Therefore, there is a need to develop a system that reduces the delay in granting leave application. The need for a web-based system is therefore based on the fact that there is a large volume of paperwork that is processed manually and takes a long time to process and that the application needs to be reduced. This does indeed cause a great deal of damage or danger to the overall productivity and efficiency of employees in the organization. There is a need to develop a system that reduces the delay encountered in the granting of leave applications.

Therefore, developing a WBLMS will help reduce the downtime as well as the paperwork and manual record-keeping information about the worker quicker with a quick analysis of the problems. The main aim of this study is to build a user-friendly web-based/online integrated public service leave management system, which is focused on a functional leave management system in the public service domain. In this paper, attention is focused only on public agencies using the African Regional Center for Space Science and Technology ducation in English (ARCSSTEE), Ile-Ife, Osun State, Nigeria as our case study.

A Web-based Leave Management System (WBLMS) is described as a Human Resource (HR) software web application that is designed to manage human capital, leave requests, procedures and records. In other words, it is seen to be a useful HR subsidiary system set aside for overseeing employee leave matters. LMS is simple to understand, easy to use and more convenient to implement.

The main aim of LMS development for the public s ector is to shift from a traditional conventional model to a more reliable and effective level of leave management (MohanaPriya, Shyamala, and Dharshini, 2017, Ehlinger, 2015). Nowadays, the current way of managing leave requests is paper-based for many public organizations in Nigeria. Nonetheless, the benefits of introducing a web-based leave management system include the easy application of leave requests, easy tracking of employees on leave, coordinated leave workflow process, reduced accumulation of leave forms, structured data presentation for report analysis generation and quick approvals. Public service organizations are known for the large human capital workforce. Leave management remains a vital area to be considered in terms of proper leave records management. The much-desired new public management system in Nigeria can only be achieved if we begin to adopt more implementation of management information systems in eliminating redundant operations.

According to the public service rule (2008), leave is the permitted absence of an officer from duty for a specified period. This act of taking time away from work cannot be dismissed or made irrelevant since every worker has the right and the opportunity to ask for time to rest or to settle personal activities outside his or her work environment. There are different forms of leave and the reasons why employees apply for leave away from work. The provision for the different forms of leave such as annual, sabbatical, sick, maternity, study, casual, pro-rata, pre-retirement and deferred leave are important to worker's physical and mental well-being. If structured properly, it can result in a desired positive impact on occupational health and safety as well as improve productivity and better performance in the workplace (Jolkkonen and Ghosheh, 2015). In this paper, the challenging issues affecting the effective and efficient management of leave demands and approvals of employees are examined to helps the Human resource managers stay up-to-date on their level of communication between the employees and their supervisors in organizations. The paper also proffers some solutions by suggesting what needs to be done to overcome the loss of time and delay in obtaining leave approvals. The aim of this study is to develop a user-friendly web-based automated public service Leave Management System for the management of employees' records. It reduces paperwork and maintains records more efficiently thus, demonstrating the use of the application for leave requests and approvals on a timely, on-demand and secure platform.

2.0 LITERATURE REVIEW

Today, the growth of organizations is driven by the availability of a large amount of information. The Human Resources office in Organizations contributes to the organizational purpose of the public service by promoting informed decision-making on employee benefits by providing accurate, timely and usable data. However, with the rise in globalization, it has become important to identify less challenging and more convenient systems to help organizations in managing their data for efficiency and efficiency. Heeks (1998) opined that the efficiency and effectiveness of activities within an organization can be improved by developing computer-based Information Systems across every functional unit. Out of the many advantages that workers have, leave is one of the most important functions carried out by the HRM.

The Human Resource Management Information System (HRMIS) which is an aspect of MIS is one of the vital digital solutions human resource personnel use in managing most of their operations. Leave Management System (LMS) is a human resource sub-system used in managing employee leave applications. Employees use LMS to request permission for leave by filling their desired dates for management approval (Singh, 2016). Other subsystems used are Payroll System, Employee Attendance System and Employee Management System like ADP, Peoplesoft (Oracle, 2019) and others to mention a few.

Given the understanding of HRMIS, it can be deduced that a web-based LMS is a Human Resource (HR) software web application that is designed to manage human capital leave requests, procedures and records. It is captured to be a useful HR subsidiary system segregated for managing employee leave matters. LMS is simple to understand, easy to use and more convenient to implement within a workspace environment. One of the many advantages of designing a Web-based LMS for the public sector is to transcend from a traditional conventional system to a more reliable and effective system (MohanaPriya *et al.*, 2017).

The Nigerian public service sector has not lived up to its expectations in terms of the development of effective ICT and MIS based applications to improve the public sector operational performance within the nation's public service. The federal ministries as the peak of the Nigerian Public service provide the administrative structures for the implementation of government programs, plans and policies. Osibanjo (2011), stated that the structure and operations of the ministries and parastatal, particularly federal make it apparent that senior managers in the course of their daily activities do engage in conventional management operations and in taking major decisions. In regards to this, there is certainly a serious need for the ministries to have in place effective MIS. Therefore, a new public management agenda that is geared towards revamping the public and civil service sectors is the adoption of MIS technologies. This management supports efficiency, effectiveness, and performance to improve the character of the civil service operation (Ojiefo, 2015). However, with the rise in globalization, it has become important to identify less challenging and more convenient systems to help organizations in managing their data for efficiency and efficiency.

The Nigeria Government requires a departure from the traditional method of administration and the urgent need for a reviewed public sector to propel the government to its quest for sustainable, socio-economic, political and technological development (Fatile and Adejuwo, 2006). The drive towards technological advancement in public service can be achieved using IS. It is therefore apparent that to carry out these functions effectively, public service organizations require IS to mitigate recurrent issues such as poor information management, inaccuracies, non-availability of data, and lack of information for monitoring, regulating, communication and in controlling public service operations (Akindele, 2006).

Soni *et al*. (2016) and Kaushik *et al*. (2017) addressed the challenge faced by students in terms of their inability to send their leave application. The authors employed a true method Group technique. The grouping method involves getting into groups and coming up with ideas, otherwise known as brainstorming. This leave management system was developed to give students the ease of submitting their leave applications online without being restricted to a manual system that uses pen and paper. To address the issue related to attendance, the authors developed a student leave management system, where the system was implemented using Java programming language and cascade Stylesheet (CS) for android. The system also helps settle issues about student attendance records. It was designed to provide the attendance status and records of a student. The outcome showed that the system developed was incomparable and effective. This application developed is aimed at developing an online LMS that is essential to an organization. In an attempt to develop a Study LMS, the authors proposed a system for eliminating the complete use of pen and paper. Consequently, the system can be used to computerize the work process of leave applications and their acceptance Hence, leave management application will minimize paperwork and keep up records in a more efficient way.

Singh *et al*. (2017) presented a leave and payroll management system to reduce manual work and to overcome the problem of time management. In an attempt to achieve the aim, the authors developed a web portal system that provides the employees of an organization with an online platform to view their leave history and apply for leave. The payroll system was implemented using VB.net as frontend and Microsoft Access 2007 SQL server 2008 as the backend. However, the system developed is desktop-based and could allow users to use it dynamically anywhere.

Priya *et al*. (2017) addressed the delay in the manual filling of leave form and waiting to get higher officials' signature. To automate the LMS, which is central to the organization activities, the authors presented a "MOBILE HRM LMS". The Mobile HRM LMS is an intranet-based application that was developed to improve the leaves segment and can be accessed throughout the organization or a department. The main aim of building the mobile HRM LMS is to reduce time spent on leave processing. The proposed system was designed using Data Flow Diagrams (DFD) and implemented using PHP programming. The result showed that there is no necessity for manual filling of leave form and wait to get approvals. The result indicated that the application developed allows the staff of the organization to view the previous leave applied by them and its ease of use thereby, reducing time.

Furthermore, Dalke *et al*. (2017) addressed the inadequacies in the use of a conventional method of staff records management in the organization. In a bid to provide information needed to manage their employee performance, the authors presented a web-based Staff Management System (SMS) to bring about transparency, organizational accountability and to allow accurate audits by creating and ensuring SMS records as solid proof as established by Griffin and Hoyle (2009) ; and maintains a repository of employment status, educational background, staff information, event records, leave credits on sick and vacation and also generate reports (Olumoye, 2013). The proposed system is a web-based application developed using the K-means clustering approach. MySQL was used as the database tool, written in PHP and developed using web technology tools such as HTML, JavaScript, CSS, JSP on a LUNA Eclipse Integrated Development Environment with the Tomcat 7.0 Server and implemented using Java programming language. The result indicated that the developed automated system helps in reducing various costs such as manpower, staff information, and work scheduling and performance analysis.

Sayali *et al*. (2017) developed a Web-based SMS in an attempt to address the inadequacies observed in the conventional method of leave management. This system is designed to store required information about all the employees working in an organization and allows officers to perform other tasks related to staff management. Staff records are very essential to organizations and it is vital to ensure there are a storage and retrieval system in place to capture relevant information of all the employees within an organization. The web-based system was a computer-based system whose database is designed using MySQL and implemented using JavaScript. Similarly, the performance analysis of the system is implemented using the L-means clustering algorithm. However, the result obtained is robust and provides an excellent performance thereby using low disk space.

Ansari *et al*. (2017) considered the time loss in applying for leave and search for employees' records. management system. They employed a development methodology called the iterative method. Similarly, the concerns about the existing system were identified and Interviews and consultations with the users of the existing system were carried out to provide quality analysis of current processes and structures of the system. The authors in a bid to addressing the challenge specified the requirements for developing the leave management system and the system was designed using Data Flow diagrams (DFD) and Entity Relationship Diagram (ERD). The deigned system was implemented using PHP for designing the user interface and SQL for designing the database. The developed system was tested and evaluated. The outcome of the system indicated that that the downtime or delay in the process of leave request and approval is reduced. The study concluded that the development of the leave management system reduces the paperwork, time expended in treating the files thereby, increasing efficiency and effectiveness.

Reddy *et al*. (2019) addressed the challenge involved in the management of information concerning staff, and the student leaves using the common manual method of leave requests and approvals. To overcome the prevailing challenges and ensure efficiency and overall performance, the authors developed an android based Leave management system by employing software methodology using Android Software kit with a user interface module. the system was developed on a three (3) tier-based web framework namely mobile handset, web server and database, also known as presentation, application, and storage respectively. The result provided an optimized solution for leave requests, approval and tracking of leaves for staff and students. The authors concluded that the Leave management system application developed to overcome the problem of manual leave application which is time-consuming. It also helped staff and students to apply for the leave through the app and get approval from higher officials as well as providing notification to higher officials upon apply of leave and the respective user are notified of their leave status.

Nevertheless, the African Regional Centre for Space Science and Technology Education in English (ARCSSTEE) implements most leave types available in the public service rule. The leave management system of ARCSSTEE is less a manual system that uses a book register to monitor the leave application process. Here, leave applications are approved by two authorities at the center: the HOD and the Director. Leave applications only get approved when they are around, else all leave applications wait till they return which results in unexpected delays in granting leave approvals. Figure 1 provides a conceptual framework for the leave process that normally occurs when an officer plans to apply for leave. The process starts with the officer and proceeds to the head of the department for recommendation. The application needs the director's approval before the officer is allowed to proceed on leave. Thereafter, the human resource department conveys the leave approval letter to the applying officer, while Figure 2.5 shows the flowchart describing the leave application process at ARCSSTEE. Yet, the major drawback of the existing systems is that they all focused on information dissipation mainly. None of the existing works explored the use of the internet as a tool for ubiquitous data creation, assess, update or deletion.

The system architecture was designed to give the ideal representation that describes the structure and views of the system is shown in Figure 2. The system was designed with a Python Programming Web Framework called Django. Django framework uses a Model View Template (MVT) framework in defining the web application structure.

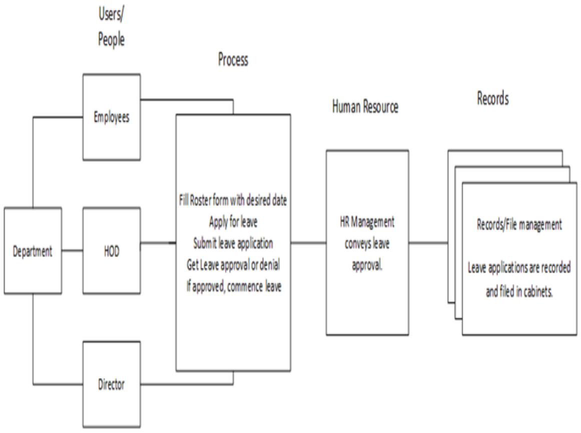

Users/
People
Process
Human Resource
Records
Employees
Department
HOD
Director
Fill Roster form with desired date
Apply for leave
Submit leave application
Get Leave approval or denial
If approved, commence leave
HR Management conveys leave approval.
Records/File management
Leave applications are recorded and filed in cabinets.

Figure1: Existing Conceptual Design Framework of ARCSSTEE leaves

The Model represents the database model, the View acts as the controller that handles all the functions and what (objects) gets viewed on the front-end and lastly the Template as the static files that are viewed by the user through the user interface. It consists of the client interface, computer software that offers service to the software application (middleware) and database. The graphical user interfaces (front side) were designed and developed using HTML, CSS, and JavaScript, while the middleware was designed based on the WGSI web server and the back-end was the SQLite database system. The web browsers present and process the web site, which is the interface for the middleware and data access layer.

The leave management system of the case-study employed the use of a qualitative research approach to gain insights and familiarity with what exist to allow further study. Interviews, scenario-generation from stakeholders and participants' experiential narratives and the analysis of current processes and structures were carried out. The system requirements were specified for the proposed system. The proposed system application requirements were gathered through the interview method with the Human Resource (HR) officer to highlight the features and specifications of the desired system. Here, ten (10) users and stakeholders of the existing system from the government agency were used as respondents. Also, ten (10) staff were selected randomly from each department were also interviewed to know what their challenges and expectations are for such a system. The requirements for the system were specified using the Use case as depicted in Table 1. The System specifications were designed using Unified Modelling Language (UML) tools (Use case, Class diagram, Activity diagram, Sequence diagram) showing the design of the proposed system as shown in Figures 3, 4, 5 and 6 respectively. SQLite database management tool was used to design and implement the database aspect of the system. The use of the database tool was motivated by its availability, accessibility, simplicity and adaptive to the implementation language tool that was used in the development of the web-based system.

Python programming language was used to implement the web-based automated leave management system application. The leave management system was implemented using web technologies, while McLean and Deleon information system (IS) theory was used to evaluate the developed system.

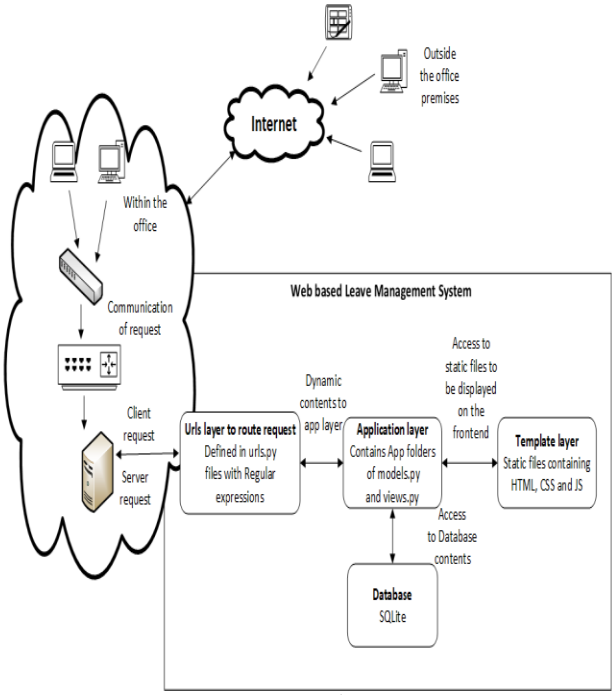


Figure 2: Overall System Architecture for the Leave Management System

Table 1: Sample Use Case Specification Table for the proposed Leave management System

S/N	Use Case name	Use Case Description	Participating Role
1	Log in.	The staff logs in to the system using a username and password to gain authorized access.	Employee, HOD, Director, HR Admin
2	Edit Profile.	The staff is allowed to update his/her profile e.g. home address, contact number, etc.	Employee, HOD
3	Request for Leave.	The staff selects a type of leave and submits his leave application for approval.	Employee, HOD
4	View Leave History.	The staff can view their previous leave applications that were approved or declined.	Employee, HOD

A qualitative method that involves a case study, face-to-face interview, and scenario-generation from stakeholder narrations were used for an exploratory purpose. About ten (10) respondents, which include the users of leave management systems (junior and experienced senior staffs) were used for data collection were gathered through interviews and scenario-generation from stakeholder narrations. Interviews provide in-depth information about participants' experiences and viewpoints on how the leave management system works.

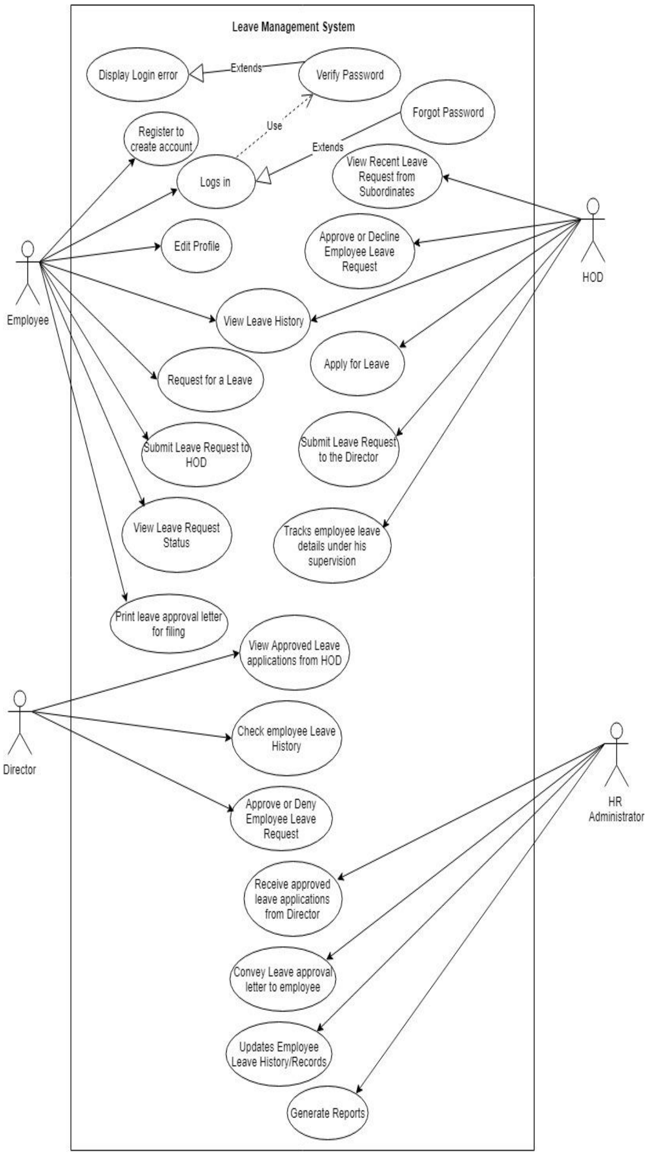


Figure 3: The Use Case Diagram Of the Proposed Leave Management System

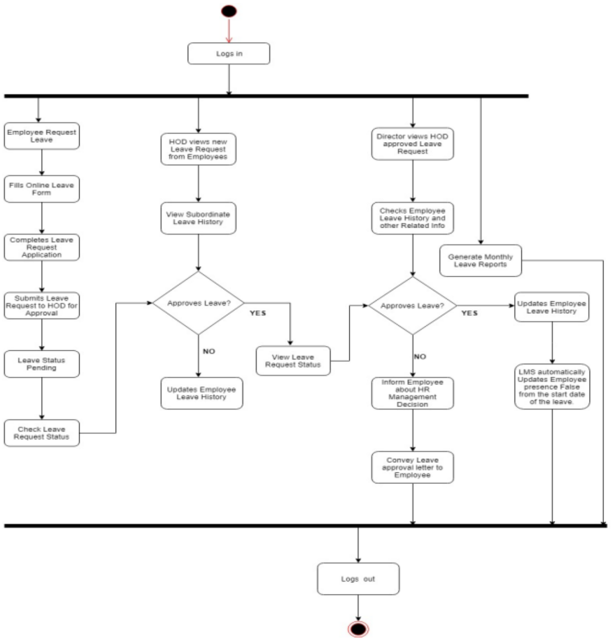


Figure 4: Class diagram for the Leave Management System

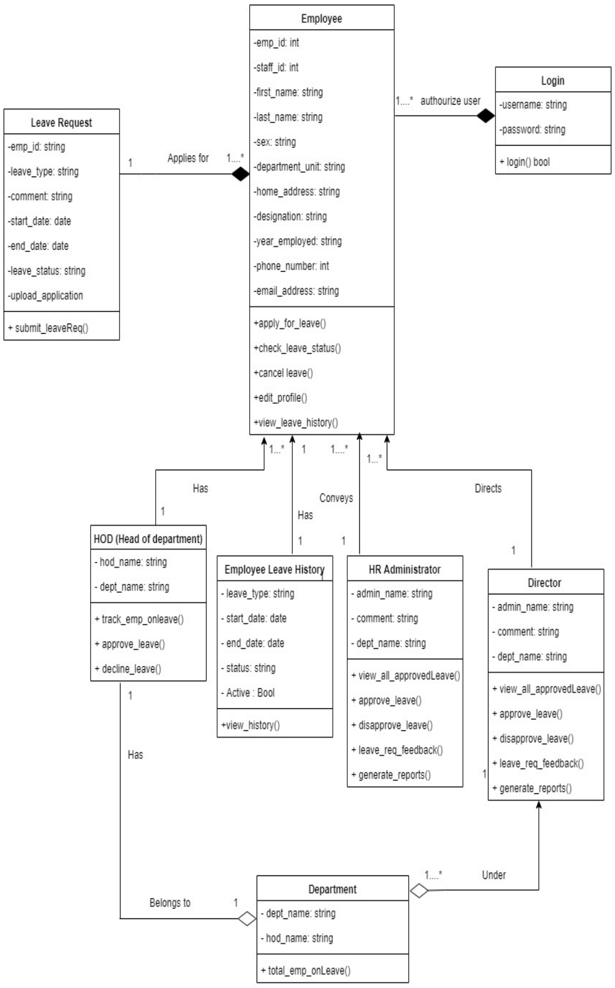


Figure 5: Sequence Diagram for the Leave Management System

This was subsequently followed by a quantitative method using a statistical approach (descriptive approach) to provide both detail and generalized analysis, which serves as the basis for result presentation. The choice of cloud service providers and users were largely governed by the intent to develop a richer understanding of how various employees' records and processes are efficiently managed for optimal performance.

RESULTS/DISCUSSION

The system was evaluated for performance based on DeLeon and McLean's information system theory. The system checks for performance accuracy and effectiveness using metrics such as information quality, service quality, user satisfaction, and net benefit through the use of the Decision Analysis Spreadsheet (DAS) tool. The DAS is a subjective evaluation analysis technique that measures the score or rate of the different categories of a parameter used in the assessment of the developed system. DAS is expressed as a single number in the range 1 to 5, where 1 refers to the lowest perceived quality and 5 is the highest perceived. The responses received from respondents are presented in Table 3.

Respondents were asked to rate the Parameters in the order of their score on the Nigeria computing system. As presented in Table 4, which is meant for substantiating each opinion and attitude of respondents relating to the quality of the developed system. The frequency of the parameters reveals the extent of the system quality as assessed by the respondents. Table 4 revealed that seven (7) respondents constituting about 70% have excellent usage with 30% of respondents have good use of the system. However, no respondent gave scores based on the other rating.

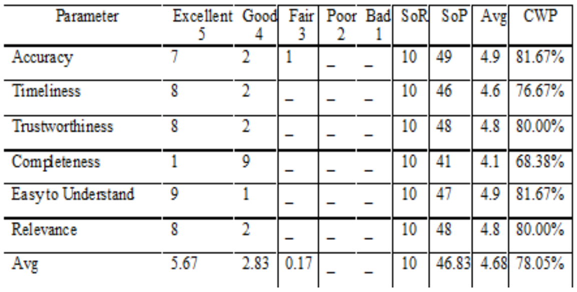
Table 3: Table of Rating Scheme

No.	Weight	Quality
1.	5	Excellent
2.	4	Good
3.	3	Poor
4.	2	Fair
5.	1	Bad

Nonetheless, 78.33% of the total respondent have the good intention of use in terms of using the Leave Management system. 73.33% of the respondents accessed the system and gave the reliability score of the system, which produces consistent results time after time. The percentage availability of 83.33% obtained from the assessment shows that the system developed is available for use. Upon the use of the system developed, the response from respondents yielded 76.67%. This is indicative of the fact that when users (junior and senior staff) use the leave system, it was considered acceptable and good enough to achieve or meet the requirement specified. Similarly, 78.33% response rate obtained from the use of the system indicated that the leave request and approval are better compared to when the manual or conventional method is applied. On average, In conclusion, the average respondent rating on the quality of the LMS produced 76.67%. this shows that the system developed was able to meet standard conditions and requirements. In the same context, the graph of the Cumulative Weighted Percentage (CWP) against the rating parameters is shown in Figure 6. The frequency of the parameters reveals the extent of the system quality as assessed by the respondents.

Based on the evaluation model, the information quality measures the content issue such that the web-based application system developed be personalized, complete, relevant, easy to understand, accurate and secured. The result obtained from the evaluation of the developed system in terms of the quality of information is shown in Table 5. The table of the score is illustrated for substantiating each opinion and attitude of respondents relating to the quality of information contained in the system developed.

Table 4: Table showing the System Evaluation Rating of the System Quality



Parameter	Excellent 5	Good 4	Fair 3	Poor 2	Bad 1	SoR	SoP	Avg	CWP
Accuracy	7	2	1	_	_	10	49	4.9	81.67%
Timeliness	8	2	_	_	_	10	46	4.6	76.67%
Trustworthiness	8	2	_	_	_	10	48	4.8	80.00%
Completeness	1	9	_	_	_	10	41	4.1	68.38%
Easyto Understand	9	1	_	_	_	10	47	4.9	81.67%
Relevance	8	2	_	_	_	10	48	4.8	80.00%
Avg	5.67	2.83	0.17	_	_	10	46.83	4.68	78.05%

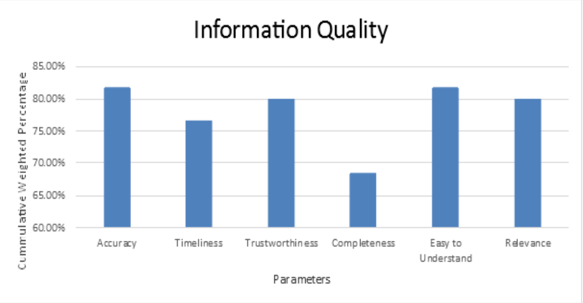


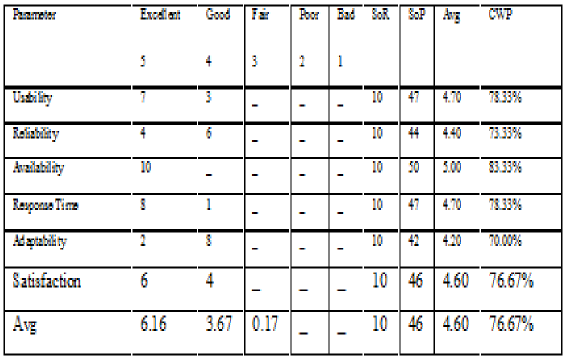
Figure 6: Graph showing the System Quality Rating

Based on the use of the developed system, the accuracy of the content (i.e. the ability to give precise outcome) gave 81.67% and having every necessary part of the requirement specified (completeness) produced 68.33% and 81.67% for ease of understanding the processes and procedures of use. Figure 7 depicts the quality rating of the developed leave management system in a graphical form.

The System Quality, Information quality singly or jointly affect the Usage and User Satisfaction of the system. The extent of use of the developed system gives 94.80% as shown in Table 6 shown, which is a process that shows the adaptation of the proposed LMS by the organization in need is excellent. However, the respondents use of the system yields a system rating of 78.05% which is also significant. This is indicative of the fact that the system quality influences usage. Consequently, the extended use of the LMS is closely related to user satisfaction as shown in Figure 6. Also, the positive experience as shown from the result obtained using the system yielded 94.80%. The high evaluation rate leads to a significant user satisfaction rating thereby, showing that the user is good with the system functionalities and operations. Similarly, the use and user satisfaction influence the net benefit in such a way that it will reduce the time used in processing leave applications thus giving a cost-effective system when compared with a paper-based leave filing system.

The System Quality, Information quality singly or jointly affect the Usage and User Satisfaction of the system. The extent of use of the developed system gives 94.80% as shown in Table 6 shown, which is a process
Table5: Table showing the System Evaluation Rating of the Information Quality that shows the adaptation of the

Table 5:



Parameter	Excellent 5	Good 4	Fair 3	Poor 2	Bad 1	SoR	SoP	Avg	CWP
Usability	7	3	_	_	_	10	47	4.70	78.33%
Reliability	4	6	_	_	_	10	44	4.40	73.33%
Availability	10	_	_	_	_	10	50	5.00	83.33%
Response Time	8	1	_	_	_	10	47	4.70	78.33%
Adaptability	2	8	_	_	_	10	42	4.20	70.00%
Satisfaction	6	4	_	_	_	10	46	4.60	76.67%
Avg	6.16	3.67	0.17	_	_	10	46	4.60	76.67%

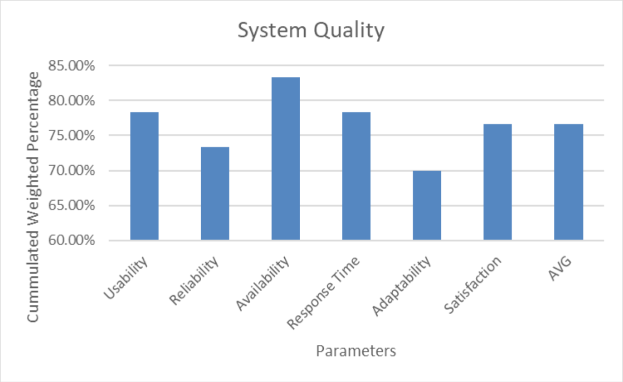
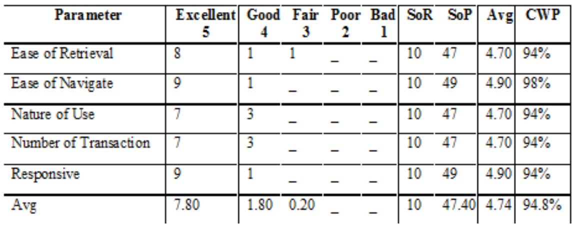


Figure 7: Graph showing the Information Quality Rating

Table 6: Table showing the System Evaluation Rating of the Usage



Parameter	Excellent 5	Good 4	Fair 3	Poor 2	Bad 1	SoR	SoP	Avg	CWP
Ease of Retrieval	8	1	1	_	_	10	47	4.70	94%
Ease of Navigate	9	1	_	_	_	10	49	4.90	98%
Nature of Use	7	3	_	_	_	10	47	4.70	94%
Number of Transaction	7	3	_	_	_	10	47	4.70	94%
Responsive	9	1	_	_	_	10	49	4.90	94%
Avg	7.80	1.80	0.20	_	_	10	47.40	4.74	94.8%

proposed LMS by the organization in need is excellent. However, the respondent’s use of the system yields a system rating of 78.05% which is also significant. This is indicative of the fact that the system quality influences usage. Consequently, the extended use of the LMS is closely related to user satisfaction as shown in Figure 6. Besides, the positive experience as shown from the result obtained using the system yielded 94.80%. The high evaluation rate leads to a significant user satisfaction rating thereby, showing that the user is good with the system functionalities and operations. Similarly, the use and user satisfaction influence the net benefit in such a way that it will reduce the time used in processing leave applications thus giving a cost-effective system when compared with the paper-based leave filing system.

5.0 CONCLUSION

In this paper, web technologies have been used to develop a web-based leave management system for the public service organization. The application of information systems for leave management systems in human resource management provides a solution to the problem of data loss, delay and downtime which is inherent in the current leave management systems. As shown in the results, web or online approaches was used for resource management in organizations and agencies. This will produce a better approach towards the management of resource, processes, and requests in public service organizations. The use of the HRMIS in form of LMS will enhance employee productivity and management communication, help the management make better and well-informed managerial decisions thereby reducing the delay in leave approval, saves time and resolve most problems encountered during the use of existing systems. The use of mail notification will greatly improve the documentation, data management and quick approval of leave requests.

When implemented, it affords the management to see the records of leaves approved for each employee or staff as well as check the leaves arrears through the World Wide Web (Web), relieving the Human Resources Department from time-consuming inquiries. It gives Human Resource administrators the ability to review, edit, and approve employee time and leave from any location by using a standard Web browser. The aspect of leave application management is a fundamental aspect of human resources and therefore, essential that delays are reduced when processing a leave application. Leave matters however can be handled more effectively if public organizations begin to appreciate the solutions that Information Technology offers as an effective supporting tool.

REFERENCES

[1] Dalke, M. S. P., Deshmukh, M. S. A., Dalave, M. J. G., Sasane, M. V. N., and Dhule, M. P. K. (2017). Web-Based Staff Management System. - International Journal of Science Technology & Engineering, (IJSTE) India. 3 (09) 271-276. Retrieved from http://www.ijste.org.

[2] DeLeon, W., McLean, E. (2003). The DeLeon and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems, 19(4), 9-30.

[3] Ehlinger, Edward P. (2015). White Paper on Paid Leave and Health. Minnesota Department of Health. Advancing Health Equity in Minnesota. Available at http://www.health.state.mn.us/divs/chs/healthequity/ahe_leg_report_020414.pdf.

[4] Griffin, A., Hoyle, M., Bastin, K., Kinyeki, C., Makhura, M., Suderman, J., ... & Wamukoya, J. (2009). Managing Personnel Records in an Electronic Environment.

[5] Kushik, V. K., Gupta, A. K., Kumar, A. and Prasad, A. (2017). Student Management System. IJARIIE, 3(5): 124-131.

[6] MohanaPriya1, A., Shyamala, G., Dharshini, R. (2017). Mobile HRM for Online Leave Management System, International Journal of Computer Science and Mobile Computing, 6 (2)17-29. Available Online at https://www.ijcsmc.com/ocs/papers/February2017/V6I2201708.pdf.

[7] Mukherjee, A. N., Bhattacharyya, S., & Bera, R. (2014). Role of information technology in human resource management of SME: A study on the use of applicant tracking system. IBMRD's Journal of Management & Research, 3(1), 1-22.

[8] Ojeifo, A., and Alegbeleye, G. I. (2015). Improving Performance of the Nigerian Civil Service through the New Public Management. Singaporean Journal of Business, Economics and Management Studies, National Library Singapore,51(2479), 1-12. Retrieved from https://*www.singaporeanjbem.com/pdfs/SG_VOL_4_(6)/4.pdf.*

[9] Olumoye, M. Y. (2013). Impact of Information Systems on Management Decision-Making in the Nigerian Insurance Sector. International Journal of Scientific & Technology Research, 2(12), 123-128.

[10] Rohilla, J. (2017). Role of Information Technology in Human Resources Management. International Journal of Advance Research, Ideas and Innovations in Technology, 3(2):566-569.

[11] Singh, M., Singh, P., Singh, R., Singh, S., Gupta, S. (2016). Leave and Payroll Management System. IOSR Journal of Computer Engineering (IOSR-JCE). International conference on computing and virtualization (ICCCV-17) PP 62-66. Retrieved from http://www.iosrjournals.org.

[12] Soni, G., Nagar, K., Fumakiya, M., Raghuvanshi, N., and Kadam, K. (2016). Study Paper on Student Leave Management Application. International Journal of Engineering Sciences & Research Technology (IJESRT) India, pp572-575. Retrieved from http: // www. ijesrt.com.

[13] Ansari, M., Qureshi, R., Halima, S. M. K., Shenaz, M.B.R., and Balashree, Y.P.R. (2017). Leave Management System. Bachelor of Engineering Dissertation Submitted to School of Computer Engineering and Technology, Anjuman-I-Islam's Kalsekar Technical Campus Mumbai, India.

[14] Reddy, E. R., Preethi, G., Murthy, H. N., Reddy, M. R. and Veena, G. (2019). Android Based Leave Management System. International Research Journal of Computer Science (IRJCS) IS SN: 2393-9842 Issue 06, Volume 6 (June 2019) www.irjcs.com SPECIAL ISSUE - 5th International Conference - "ACCE – 2019"

THE ROLE OF ICT TOOLS FOR ECONOMIC DEVELOPMENT AND POVERTY REDUCTION

Nwakpa Matthew Nwankwo
Department of Computer Science, Ebonyi State College of Education, Ikwo, Ebonyi State, Nigeria
monnon200@gmail.com

ABSTRACT

Various ICT tools can be used for business transaction to aid economic development of the Nation. This various ICT tools have positive implications for the growth of businesses as there are many participant in its operation and that will also benefit from it. Despite the benefits that can be driven from using these tools for poverty reduction, their roles are still very limited. This paper succinctly examines these tools and their roles that will enhance ICT development

implementation and poverty reduction. It highlighted the implications of these tools as it is always dynamic in nature for its application in Nigeria business program. Furthermore, the paper discusses the prospect of these tools needed as users needed continuous training/usage of those tools. Its provision is a must for this effort to be achieved which must include needed manpower to manage and enhance their performance operation. In addition, It made recommendation on how to achieve and make this effort functional and tries to concluded by saying that those tools is needed and indispensable in business operation. The insight obtained from this research can be of great benefits for economic decision-making and poverty reduction by government agencies.

Keyword: e-learning, Information and Communication Technology (ICT), ICT development, poverty reduction, ICT-based business

1.0 INTRODUCTION

At this present age, information and knowledge have become strategic and essential tools that would aid nations to join or make the world a global economy. Moreover, informed citizen, government, public sector can make their decision at every time. Moreover, economic and social development of our nation can be accelerated through adoption and assimilation ICT tools. However, the population of most Nigerians that have utilize ICT infrastructure for economic growth are still very low [1]. Information and communication technology (ICT) tools can provide positive impact on the social economic development of individual and the society at large. Consequently, if nations are to build and organize self-reliant technologically compliant society, the need for development of advanced ICT tools cannot be overemphasized. And Nigeria requires this tools in areas such as political development, educational attainments etc. [1]. Typical examples of ICT tools for that have played invaluable roles in economic development and poverty reduction include but not limited to Facebooks, Instagram, different Google tools, Twitter etc.

The effect has made Nigerians participation in ICT development process very low. This ICT tools where available can have a positive impact on people's lives and equally on the socio-economic development of our nation. At present this tools have affect in every areas of lives, people are empowered to manage and disseminate information as in governance, Education, health Agriculture, disaster management peace-building economic and social integration. According to Okoh [2], with appropriate ICT tools, citizens can change the equilibrium of power, make informed choices and create opportunities on their lives. A consideration of this enormous potential of ICT will help Nigeria to develop and join the world at large in globalization process or global information society. In recent research, Agba [3] noted that the major area in which ICT tools have greatly affected peoples' lives include educational systems, commerce and industry, manufacturing and social system.

ICT tools in the global environment has provided unparalleled facilities and opportunities for any Nation to join and enjoy it's benefit with ICT fully utilized citizen can work at the comfort of their home thereby reducing unemployment opportunities. People can work through the internet. With internet fully in place of work in the country, people can work remotely through virtual office. Contacts and business can be transacted through the internet. With the introduction and utilization of the mobile phone in the area of internet, this enormous potential are so much as it can be used to create web site for individual business or corporate organization. Moreover, it can attract foreign investment to our nation [5].

The major issues that has hampered the utilization of potentials of internet for economic development is that, the people in more remote or agrarian areas are sometimes unaware of the importance of this ICT tools. With ICT in Education, teaching and learning can be transacted for people that do not undergo formal education which has no age limits. In addition, ICT tools has importance in areas such as e-learning, e-book, e-library, e-marketing.

The major advantages ICT tools in the educational learning is that it has no age limit for learning. However, factors that can be considered in this approaches are technical knowhow, infrastructural resources such as computer and the patient to learn the enormous resources that internet provides to enhance economic wellbeing [6].

2.0 LITERATURE REVIEW

2.1 Issues Hampering Adoption of ICT Tools for Economic Development and Poverty Reduction

The e-transactions/ICT tools has being fully in places in most part of the country but most people's perception towards the system is not encouraging. Majority refuse to fully use it to conduct business simply because they fill that the network will fail them. Majority of the users are seen as being corrupt, fraudulent and dubious. This stems from the fact that few of the users/fraudsters are using the means to dupe their foreign counterpart that makes them bear a common name in Nigeria called *Yahoo Yahoo*. The developing country such as Nigeria and other African countries flood the internet with non-existing companies and facilities [7, 8]. Internet accessibility is no longer the problem due to the involvement of telecommunication companies such as MTN, Glo, Etisalat etc. in the internet service delivery. People do not fully utilize the potentials posed by this system for jobs participation due to fear or lack of proper knowledge. Thus majority of the population of Nigerians are not interested thereby leaving most people still live in abject poverty. Most people do not know that they have oil well with their hand phone or better laptop with other ICT gadgets/tools. Success in ICT tools require investment in time and money [4] success doesn't happen overnight, never stop learning, asking question and trying new things.

Other issues that limit full utilization of ICT tools for economic development and poverty reduction are also highlighted below. The reason most people fail to make money on the internet is because they don't have the knowledge and tools they need to market their business effectively and successfully. They just end up spinning their wheels because they don't know where to start and neglect to develop and follow a consistent market plan.

A blueprint is vital when you are first starting out. If you don't have a step by step plan knowing where you are going it is like a car without a roadmap. A step by step plan takes you by the hand showing you what needs to be done through daily tasks.

Set aside so many hours daily in the week as to when you are going to work on your business plan and stick with it. Consistency is the key. Treat your business like a hobby and you will get mediocre results. If you treat your business seriously expect to accomplish a better result. There is no free ride, realize that and you are a head of the game. Stay consistent, use the best tools and treat your business seriously. Find a mentor, support person or even a support network. This is vital when you are first starting out.

This support person can share the wealth of their knowledge telling you what works and what doesn't work, and they can keep you motivated. You can price a great deal of free information and have many questions answered [6].

Okoro [4] success requires an investment in time and money. Bottom line is that there is no free ride; understand this and you are on your way to success. Above all be patient and never give up. Through persistence, support, using the best tools and believing in you. A good foundation is the key and then consistency working on it, produces result. Success doesn't happen overnight. Never stop learning, asking question and trying new things.

The simple fact is that so many of the get rich in the internet are scamming. Many people place themselves as being rich on the internet but actually they are not successful. Many of the successful program they have are not actually true.

In-fact their only income is generated by scamming new comers. Always believe in yourself that nobody gets rich quick on the internet but millions are there for the making if only you put in your time and learn. Always practice these three key words which include study, learn and ask question.

Don't sign up for any scheme, program or partnership that involves sharing your credit card information until you have thoroughly studied the situation from all angles. Don't let all that is ugly and bad overshadow the true good that can be accomplished on the internet.

Most study centers on the internet claim to award certificate for a course of study that takes years in a matter of weeks. Avoid such school or study programme as they can only collect your money and provide you with nothing. Stick to schools that give proper training and guidance. There is also the problem of Hackers and fraudster that exist whose mission is to temper with people's data.

Quality internet education is a fundamental requirement in the school of internet success. There is no shortcut to internet success. Study the works and life styles of successful internet entrepreneurs, follow their footsteps and in no time you will be the next. There is common saying that nothing good comes so easy to achieve.

People should keep away from internet scams and other activities that constitute cybercrime. It is time we understand that for every scam we commit online, we are taking the image of our dear country backward on the opportunities that abound on the internet.

No private sector pays as much as their government counterpart be it (university, secondary school, primary, companies, and business sector). The reason is not farfetched because the private sector pays from the money they realized from the company. Government should look inwards assuming the oil dries up or have less value or non-value what will happen to the masses that wholly depend on government white cola job. How it's employee to be much productive in their sit down offices and discuss their family or other matters waiting for the month to end. In the private sector it is not so it is either you work or will not be paid. The students in private schools pays much school fees more that the public school yet their teachers are paid less.

2.2 Areas that have benefited from ICT Tools to Improve Economic Development

There are various areas that have seen high impact research on the role of ICT tools for economic development [1]. In this section, these areas are outlined with their effective strategies for improvement.

e-Learning (or electronic learning): This covers every technology based learning (TEL). Any learning carried out online system or web based learning. E-learning is learning which makes learner attend a course of study or programme where the students rarely or never meet face to face, no access on campus/environmental educational facilities because they study online typical example is the National Open University.

Children and adult learning pattern is improved upon when they are exposed to the latest information or current thing. ICT enhances learning pattern of people, most people can learn faster if ICT is in place. If anyone have the financial muscle he can buy any current book which he or she can subscribe for called e-book.

e-Business – The good thing about online business is some many opportunities. You need to learn and takes time to learn how to do business on the internet and it also takes time to be success on the internet. In internet business you must explore your opportunities. Remember that the internet really is the information super highway. Okoro [4] study, learn and question continue to do all the three even after you have started your own business and even after you have made your money. Some of the known activities done of the internet are

Web-hosting services, Web site design and consultation, Marketing advice, Traffic, Education and training, Advertising services, Affiliate programs,

Computer hardware and software, Credit card services, others online business programme that generate income are pay to read, High yield investment programs, survey/data entry companies, forex, trading, google adwords and adsense so many of them [9].

3.0 HOW TO GET PEOPLE’S INTEREST IN USAGE ICT TOOLS FOR ECONOMIC DEVELOPMENT

Government can be involved in that aspect. Majority of Nigerians depend on government to employ and pay them monthly. Government can create ministry of ICT with different departments that do online business like department of FOREX Trading, department to use Google Adwords and Adsense, survey/data entry companies department, etc or create Agency for ICT implementation or ICT development board. They can also have functional ICT department in each ministry that fully participates in this online business.

In that process they monitor that good and the bad in the system. The oil we sole depend on will not be much source of energy in the developing world of ours or the oil may finish. Government as the driving force that the masses look up to there is need to exploit that sector. There is a saying that anywhere the child is crying and pointing at either the mother is there or the father.

Since young Nigerians that engage in this online business are making so much money becoming so rich. There is need for government to embark on this sector and employ masses to practice them; this will encourage private partnership to join that line and remove or eradicate extreme poverty in the land. It can greatly enhance the achievement of millennium development goals if applied appropriately without hurting anybody.

To eradicate this extreme poverty, hunger and unemployment in the land they must be political will change. People can be poor due to lack of opportunities, ICT sector can be harnessed to make jobs available for the citizens government should create this online jobs that will yield something for the government and the citizens monitor it and give them target instead allowing people stay idle for a whole day inside where they call offices.

The problem with us is lack of good governance and transparency. The policy may be good but with corrupt mind implementation becomes the issue that can add to frustration. With ICT in place, this can help in the area of information sharing government participation in ICT will help a lot to even connect the rural and urban areas to enable information to flow this has been achieved lately due to introduction of GSM mobile phone.

Okoh [2], a well-informed society will be ready to eradicate hunger because when people are aware of where to get held, they usually engage in it. Most of this online business can connect Nigeria and exploit what we do and earn a living. Farmers in the rural areas need to know what their counterpart are doing in advance countries and how they too can improve their harvest.

The only difference is that it will take time, effort and investment to make it online.

Look at the newly created Peace Corps, 2015-2018 where most Nigerians after registering where disbanded due to non-approval by the National Assembly. Government can do the same for the ICT sector, you can call it anything, with this in place the fraudsters in the system can be detected, arrested and jailed. Thus government becomes the manager of that sector employ many Nigerians that will be ready to generate income for the government thereby pushing the economy to the better level.

4.0 SUGGESTIONS TO IMPROVE ICT TOOLS ADOPTION

1. Government at all levels should embrace ICT for their financial transactions e-government. The government should create ministry of ICT or ICT development board at all levels.
2. The government should employ the masses/youths in that sector and place them on probation just like the N-power programme created by the present administration any individual that is ready to work and generate income will automatically be observed in the nation’s civil service.
3. The government can have ICT training school for internet business for the youth empowerment and make it affordable or free just like Nigeria air force school and others.

4. The culprits that dent our image should be dealt with in a decisive way to serve as a deterrent to others.
5. Youths should channel their energies at identifying and using aspect of ICT for business that would be beneficial to the country at large.

5.0 CONCLUSION

ICT usage especially in business transaction deserves better attention as it can create opportunities that will push the nation's economy to a greater height; since ICT can be applied in all aspect of life resulting in better services. It can lead to the achievement of millennium development goals which are purely developmental and as a tool for integration of nations to form a common platform at tackling problems of unemployment, underdevelopment and other benefits from economics of scale.

However such glowing attributes to use ICT in business does not mean there are no challenges. The challenges are much as people's feeling, capacity building and misuse of service by malicious individuals.

It is worthy to note that how well our youths are able to develop and take advantage of the opportunities available in ICT sector would determine our quest for development and integration in the committee of Nations. Adequate steps have to be taken to ensure that the youths are empowered and encouraged to use the system. The Nation will be at a better forefront to fight poverty, hunger, underdevelopment and other opportunities that we are still missing out today in the world.

You need to rework this paper to provide empirical evidence from literature, and to show possible survey you have conducted in this regard. The ideas in the paper are good, but it lacks evidence and novelty.

REFERENCES

[1] Nwabueze, A. U., Ozioko, R. E. (2011). Information and Communication Technology for sustainable Development in Nigeria. *Library Philosophy and Practice.*

[2] Okoh C. (2013) – *Unpublished paper on ICT* Nigeria computer society conference, Abakaliki.

[3] Agba P. C. (2001). Electronic Reporting: Heart of the New Communication age. Nsukka University of Nigeria Press; 26-30.

[4] Okoro Victor I. (2008) – *Internet manual July*, Edition p. 5.

[5] Johnson A. (2008) *Utilization of Information and communication technology (ICT) P.* 122-123 Enugu.

[6] Wyun G. (2007) *Impact of information and communication technology ICT* p. 40 Hippuks additional press Enugu.

[7] Barnat C. (2006) *Cyber business mindset for a wired age*-John Wiley and Sons Ltd Englan.

[8] Bernstein Terry Bhimani, BA Schultz and Siegel AC (2006). *Internet Security for Business* – John Wiley, New York.

[9] Elshadai J (2006) – *E-commerce Break through Manual* – volume 1.

AUTHOR'S BIOGRAPHY

Nwakpa Matthew Nwankwo received his Master of Science (Computer Science) from Ebonyi State University, Nigeria and Bachelor of Engineering (Computer Science Engineering) from Enugu State University of Science and technology (ESUT), Enugu, Nigeria. He obtained his PGDE from Usman Danfodio University, Sokoto, Nigeria. He is currently Lecturer I at Ebonyi State College of Education, Ikwo, Ebonyi State. His research interests are ICT tools for economic development and poverty reduction, educational and effective learning.

A MACHINE LEARNING APPROACH FOR PREDICTING DRUG REACTIONS FROM PATIENTS' CASE NOTES

*[1]Agbonkhese C. and [2]Soriyan H. A.
[1]Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria
[2]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria

*Email of Corresponding Author: agbonschris@gmail.com

ABSTRACT

Timely and highly accurate identification of potential adverse drug reactions (ADRs) emanating from drug prescriptions is very essential for quality healthcare delivery for public health. Existing computational techniques using machine learning algorithms for detecting drug reactions most rely either on the chemical structures of the drugs or on genomic information, using 'guilt by association rule'. Recent studies have adopted the combination of both the chemical structures of drugs and gene expression variables. However, due to individual variations in response to prescribed drug, many patients still suffer drug reactions from drugs, even after proper screening during the drug discovering process. This paper attempts to present a novel approach for proper prediction of drug reactions emanating from drug prescriptions during patients care, using clinical data of patients such as age, weigh, diagnosis, prescribed drug, route of intake, Polypharmacy, alcohol dependent, and drug dependent. The data was collected through record overview of seven hundred and twenty four patients' case notes, who were admitted within the last five years at the mental health department of the Obafemi Awolowo University Teaching Hospitals Complex, Ile-Ife, Nigeria. Data relating to these variables was used as input for training a Gaussian Naïve Bayes supervised machine learning algorithm in python programming language which predicts the binary class label; drug reaction and no drug reaction. The model had an overall accuracy of 64%, with an error rate of 0.36. In general, the accuracy of the model was significantly good enough to predict drug reactions in a clinical setting, with a considerably low error rate.

Keywords: drug reaction, drug, prediction, machine learning, patient.

1.0 INTRODUCTION

In a typical drug discovery cycle, from the point of target identification to the point of clinical use can take approximately 14 years [1], incurring an associated cost of about 2.6 billion dollars [2]. One of the main reasons while drugs fail in the process of developing drug is the existence of Adverse Drug Reactions (ADRs). ADR is universally known as a serious clinical problems and it is estimated to cause far more than two million hospitalizations [3] and over one hundred thousand deaths in the United States alone every year [4]. Also, in case of sever ADR, pharmaceutical companies would be forced to withdraw their drugs from the market. This involves significant danger for the patients, as well as major financial implications to the various companies involved in the drug production. Based on this, it became imperative to predict ADRs prior to market introduction of the drug, and has been considered as a very challenging issue in drug development process. Using laboratory-based approaches for ADR prediction and evaluation of its potential are often very costly and time consuming. Therefore, the use of computational approaches for early detection of potential ADRs has gained much attention in the recent years. The general style for computational approach is as follows: Firstly, they consider several different chemical and biological properties of the drugs. Secondly, they transform the considered chemical and biological properties into Numerical features. Thirdly, they then develop a systematic way of comprehending, predicting and finally interpreting the desired and undesired effects of the drugs [5-8]. The fundamental difference that exist among these various methods lies in the type of properties they consider for the ADRs analysis. In the most research domain, more fascinating knowledge can be mined from the relationships among variables [8 & 9]. For instance, several studies have shown that considering the relationships that exist among different diseases reveals very useful informative patterns and is useful for different prediction demands [10]. Most of these existing methods focus on the drugs individually, and neglecting the informative knowledge that could be derived from the hidden relationships among different drugs and the characteristics of the features of the individual patients.

There are some recent approaches that follow the assumption that 'similar drugs have similar adverse effect' pattern and consider the relationships among drugs for predicting ADRs [11].

In this paper, we consider information hidden in the relationships among drugs as well as that of the patients for which the drugs have been prescribed, and apply the Gaussian naïve Bayesian network classification algorithm to predict drug reaction.

2.0 RELATED WORK

[12] reveals the relationship between drugs and their side effects using side effect resources from SIDER and biological process information from GO. Their experiments were done to prove an assumption that phenotype can be impacted by action between drugs and biological processes. Accordingly, two essential networks, drug biological process network and drug side effect network were merged to derive the multi-level side effect related biological processes network. The authors in [13] adopts the ligand-based

method for predicting drug reaction, which presumes that if two drugs have a similar chemical structure, they usually have similar drug side effect. The advantage in this approach is simply to infer the drug side effect of another drug by manipulating chemical properties with known drug side effect [14]. However, the chemical structure cannot provide any biological interpretations regarding the underlying mechanisms at a molecular interaction level [15]. Moreover, sometimes they are different chemical structures but trigger the same drug side effect. The authors in [16] used correlation-based approach for predicting drug reaction. The main process of their method is to calculate similarity score between drugs and their side effects. In the first step, the method calculates and ranks the correlation scores. Consequently, the potential score is derived to make a decision on the side effect given drug.

[17] proposed a novel framework to predict ADRs of a given drug by considering on canonical correlation including network-based diffusion. The work could significantly improve prediction performance from CCA approach. A recent effective method to predict ADRs and therapeutics is similarity ensemble approach (SEA). This is motivated by two concepts; the proteins can be related by their pharmacology and the established drug by exploiting the biological networks can exploit the relationship of side effects or protein targets. The authors in [18] used ligand-based information from ChEMBL to examine the similarity using SEA. The closest similarity score between two drugs implied that they may occupy a same side effect.

[19] proposed the ‘promiscuous’ to encompass data source of drug-side effect, drug-target, and protein-protein relations. The structural similarity of drugs was applied to discover relations among proteins, side effects and drugs. Although it could provide efficient network visualization, drug-target interactions have to be validated and tested experimentally due to their complexity of interaction characteristic.

Recently, machine learning approach has been considered using computational methods. Computational methods and text mining takes into account to analyze ADRs relationships and generate predictive models. The author in [20] proposed the utilization of chemical structures and biological properties or phenotypic characteristics utilizing four data sources to create high dimensional feature vectors. The encoded fingerprints corresponding to chemical structures were obtained from PubChem. The biological properties, targeted proteins, transporters, enzymes, and protein pathways, were collected from DrugBank and KEGG. The well-known drug side effect database, SIDER, provided the phenotypic of drugs, such as treatment indicators and other side effects. To assess the effectiveness of the predictive model, five well-known algorithms namely logistic regression, Naıve Bayes, k-nearest neighbor, random forest, and SVM were employed. The results suggested phenotypic data were the most beneficial to predict ADRs. The author in [21] however combined phenotypic data with baseline biological and chemical structures, and achieved a predictive model with signification improvements.

3.0 METHOD

3.1 Data Collection

Data was collected from the mental health department of the Obafemi Awolowo University Hospitals Complex (OAUTHC), Ile-Ife. A total of seven hundred and twenty four (724) data of patients who were admitted within the last five years was collected, and a total of thirty one (classes) were found within the variables considered. The variables include: sex, age, weight, diagnosis, Polypharmacy, alcohol dependent, drug dependent, breastfeeding, pregnancy, prescribed antipsychotic drugs, route of intake, other drugs, and adverse reaction.

3.2 Data Preprocessing

As part of the preprocessing, all Boolean equivalent of ‘NO’ as well as missing values were replaced with zeros (0), while all binary equivalents of ‘YES’ were replaced with ones (1). Feature selection was manually done by excluding features with constant values of zeros (0), such as that of breastfeeding, pregnancy, alcohol dependent and drug dependent selected out of the dataset. The dataset initially had an total of thirty four (31) features, and after the feature selection process, the total features was reduced to twenty seven (27)

3.2 Further Classes

Some of the variables whose values were not Boolean values were further categorized, these variables include; diagnosis, prescribed antipsychotic drugs, and route of intake.

3.3 The Diagnosis

The diagnosis of the patients include; Psychotic disorder, Schrizophrenia, Bipolar addictive disorder, Substance/drug abuse, Organic psychotic disorder, Mental and behavior disorder, Sever depression, Mania, Somatic disorder.

3.4 Antipsychotic Drug:

Haloperidol, Diazepan, Chlopromazine, Trifluoperazine, Fluoxetin, Risperidone, Olanzepine, Benzhexol, Flupentixol, Fluphenazine, Flupentixol, Amitriptyline, Carbamazepine, Hariprazine, and Bromazepam.

Route of intake: Oral, intramuscular, and intravenous.

Naïve Bayes Model

Bayesian theorem deals with strong independence assumptions between predictors. The theory provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). The classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. That is, to say that the various attributes that predict the possibility of an adverse drug reactions are independent of the others. The mathematical equation for Naïve Bayes is given as follows:

$$P\frac{C}{X} = \frac{P(X \mid C)P(C)}{P(X)} \tag{1}$$

Where:

P(c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

Let D be a training set of tuples and their association class labels. As usual, each tuple is represented by an n-dimensional attribute vector, *X* =$(x\_1$,x_2 ,…, x_n), depicting *n* measurement made on the tuple from *n* attributes, respectively, $(A\_1$,A_2, … ,A_n).

Suppose there are m classes represented as$(C_1$,C_2, … ,C_m), given a tuple, X, the classifier will predict that tuple X belongs to the class C_i if and only if:

P (C_i |x) > P (C_j |x) for 1 <= j <= m; j ≠ I (2)

Thus this maximize P(C_i |x).

The class C_i for which P (C_i |x) is maximized represents the Maximum Posteriori Hypothesis.

By Bayes' theorem

P (C_i |x) = P ($x|C\_i$) P($C\_i)$ /P(x) (3)

As P (x) is constant for all classes, only P ($x|C\_i$)P ($C\_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is;

P (C_1) = P (C_2) = ………… = P (C_m) (4)

Maximizing P($x|C\_i$). Otherwise, this maximize P($x|C\_i$)P(C_1). That the class prior probabilities may be estimated by:

P (C_1) = |C_1, D|/|D| (5)

Where:

|C_1, D| is the number of training tuples of class C_1 in D.

Given the above kind of datasets with many attributes, it would be extremely computationally expensive to compute P ($x|C\_i$). In order to reduce computation in evaluating P ($x|C\_i$), the naïve Bayes' assumption of class conditional independence is made.

As per the conditional independence assumption of Bayes theorem, the presence or absence of some parameters of a class is independent to the presence or absence of some other parameters, making each parameter's contribution independent to the final result. For instance, for a parameter P(ADR = "Yes") given "polypharmacy" = 'Value from Test Data' is independent of P(ADR = 'No') gives "polypharmacy" = 'Value from Test Data'. In similar way, that is how the probabilities of all the parameters and their individual contribution to the final result in different variables could be calculated.

3.5 Framework of the Proposed Model

Figure 1 describes the proposed framework for predicting adverse reactions from drug prescription, using the dataset derived from patients' case notes. These data was first of all subject to manual feature selection process, and seventy percent (70%) of the dataset relating to the selected features was used to train a Gaussian Naïve Bayes classifies which works with idea of conditional independence derived from Bayes theory, while the remaining thirty percent (30%) of the data was used to test the model. After which the results derived was evaluated.

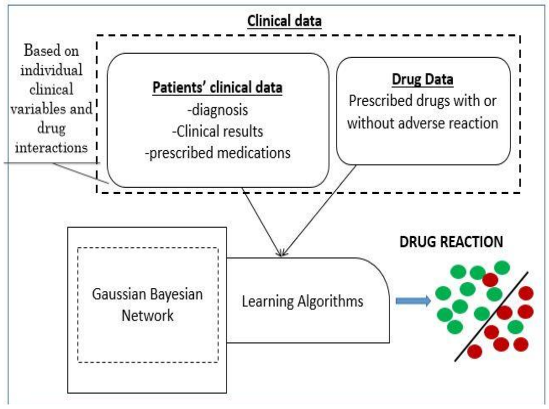
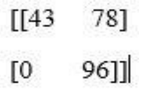

Clinical data
Based on individual clinical variables and drug interactions
Patients' clinical data
-diagnosis
-Clinical results
-prescribed medications
Drug Data
Prescribed drugs with or without adverse reaction
Gaussian Bayesian Network
Learning Algorithms
DRUG REACTION

Figure 1: Framework of the ADR prediction model

4.0 RESULTS

The summary of the results of the prediction model is as given in Table 1, which was also systematically derived from the confusion matrix in Figure 2.



```
[[43   78]

 [0    96]]
```

Figure 2: Confusion matrix of the prediction model for drug reaction

The model had an overall performance accuracy of 64%, with an error rate of 0.36. From the confusion matrix, out of a total of two hundred and seventeen (217) instances in the dataset used for testing the model, a total of ninety six (96) cases of diagnosis and corresponding drug prescriptions were correctly predicted to result in adverse reactions, producing a precision of 0.55. A total of seventy eight (78) instances were wrongly classified to resulting in adverse reactions. A total of forty three (43) instances were correctly predicted to not resulting in adverse reactions, while there were no instances of predicting that a patient would not have an adverse reactions when he/she actually would have. The prevalence, which characterized the 'yes' condition or actual occurrence adverse reactions in our sample is given as the average weight 80.

Table 1: Summary of the results of the performance of the prediction model for drug reaction

	Precision	Recall	F1-score	support
0	1.00	0.36	0.52	121
1	0.55	1.00	0.71	96
accuracy			0.64	217
Macro avg	0.78	0.68	0.62	217
Weighted avg	0.80	0.64	0.61	217

5.0 CONCLUSION AND FUTURE WORK

In this paper, we present and discussed recent approaches for predicting drug reactions and describes a machine learning approach that can be used for predicting adverse drug reactions. In doing this, we present the description of adverse drug reaction prediction model using the Gaussian Naive Bayesian classifier available in Python programming language library, trained and tested with the data collected from patients' case notes from the mental health unit of the Obafemi Awolowo University, Ile-Ife, Nigeria. The performance of the model was significantly good enough in predicting drug reactions. The model we have presented here can be further implemented and integrated into a hospital information system that manages drug dispensing such as the 'E' version of the Made in Nigeria Primary Healthcare Information System (E-MINPHIS) this will greatly alleviate drug reactions emanating from drug prescriptions during patient care in a clinical setting. However, in future we wish to integrate different data sources and using other machine learning algorithms to improve the performance of the prediction model for drug reactions.

REFERENCES

[1] Myers, S. Baker, A. (2001). Drug discovery; an operating model for a new era, *Nat. Biotechnol. 19* 727–730.

[2] DiMasi, J.A. Hansen, R.W. Grabowski, H.G. (2016). Innovation in the pharmaceutical industry: New estimates of R & D costs, *Journal of Health Economics* 47; 20–33

[3] L.J, P. BH, C. PN, (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies, *J. Am. Med.Assoc*. Q5 279(15) 1200–1205,

[4] Giacomini K.M., Krauss, R.M. Roden, D.M. Eichelbaum, M., Hayden, M.R Nakamura, Y. (2007). When good drugs go bad, *Nature 446* (7139) 975–977

[5] Hammann, F. Gutmann, H. Vogt, N. Helma, C. Drewe, J. (2010). Prediction of adverse drug reactions using decision tree modeling, *Clin.Pharmacol.*Ther.88 (1) 52–59

[6] Xie, L. Li, J. Xie, L. Bourne, P. E. (2009). Drug discovery using chemical systems biology: identification of the protein–ligand binding network to explain the side effects of cetp inhibitors, *PLoSComput*. Biol.5 (5) e1000387, http://dx.doi.org/ 10.1371/journal.pcbi.1000387.

[7] Yamanishi, Y. Pauwels, E. Kotera, M. (2012). Drugside-effect prediction based on the integration of chemical and biological paces, *J. Chem. Inf. Model*. 52 (12)3284–3292,

[8] Fukuzaki, M. Seki, M. Kashima, H. Sese, J. (2009). Side effect prediction using cooperative path ways, in: Proceedings of the 2009 IEEE *International Conference on Bioinformatics and Biomedicine*, Washington, DC, USA, November1–4, 2009, IEEE Computer Society, pp.142–147

[9] Cook, D.J. Holder, L.B. (2006). *Mining Graph Data, John Wiley and Sons,* 2006.

[10] Zhang, M. Zhu, C. Jacomy, A. Lu, L.J. Jegga, A.G. (2011). The orphan disease networks. *J.Hum. Genet*. 88 (6) 755–766

[11] Luo, H., Chen, J. Shi, L., Mikailov, M., Zhu, H., Wang, K., He, L., Yang, L., (2011) DRAR-CPI: a server of identifying drug repositioning potential and adverse drug reactions via the chemical-protein interaction, *Nucleic AcidsRes*. (39) 492 - 498

[12] Lee, S., Lee, K. H. Song, M. and Lee. D. (2010). Building the process-drug side effect network to discover the relationship between biological process and side effects. *Bioinformatics,* 12(Suppl2):S2.

[13] Scheiber, J., Jenkins, I. L., Sukuru, S. C. K. Bender, A. Mikhailov, D. Milik, M. and Azzaoui, K. S. Whitebread, J. Hamon, L. Urban, M. Glick, and J. W. Davies. (2009). Mapping adverse drug reactions in chemical space. *J. Med. Chem*., pages 3103–3107.

[14] Atias, N. and Sharan, R. (2011). An algorithmic framework for predicting side effects of drugs. Journal of computational biology: *journal of computational molecular cell biology,* 18(3), 207-218

[15] Mizutani, S., Pauwels, E., Stoven, V. Goto, S., and Yamanishi, Y. (2012.). Relating drug-protein interaction network with drug side effects. *Bioinformatics*, 28:522–528.

[16] Tatonetti, N. P., Fernald, G. H., and Altman. R. B. (2012). A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *J Am Med Inform Assoc,* 19:79– 85.

[17] Lounkine, E. Keiser, M. J. Whitebread, S. Mikhailov D., Hamon J., Jenkins, J. L. Lavan, P. Weber, E. Doak, A. K.. Shoichet, B. K and Urban. L. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *NATURE*, 486:361–368.

[18] Keiser M.J., Roth B.L., Armbruster B.N., Ernsberger P., Irwin J.J., and Shoichet B.K. (2017). *Relating protein pharmacology by ligand chemistry.* 25(2):197–206.

[19] von Eichborn, J. Murgueitio, M. S Dunkel, M. Koerner, S. Bourne, P. E. and Preissner. R. (2011). Promiscuous: a database for network-based drug-repositioning. *Nucleic Acids Res,* 39 (Database issue):D1060D1066.

[20] Liu, M. Wu, Y. Chen, Y. Sun, J. Zhao, Z., Chen, X., Matheny, M. E., and Xu. H. (2012). Largescale prediction of adverse drug reactions using chemical, biological, phenotypic properties of drugs. J Am Med Inform Assoc, 19:e28–35.

[21] Huang, L. C., Wu, X and Chen, J. Y. (2011). Predicting adverse side effects of drugs. BMC *Genomics,* 12 (Suppl 5):S11.

DISTANCE-DISTRIBUTED ENERGY EFFICIENT CLUSTERING (D-DEEC) ROUTING PROTOCOL

[1]Jibreel F., [2]Daabo M. I. and [3]Gbolagade K. A.
[1]Department of Computer Science, Tamale Technical University, Ghana.
[2]Department of Computer Science, University for Development Studies, Ghana.
[3]Department of Computer Science, College of Information & Communication Technology,
Kwara State University, Nigeria.

Email of Corresponding Author: *jibreel.fuseini@gmail.com*

ABSTRACT

Many cluster-based routing protocols have been proposed based on heterogeneity in recent times. Indeed, these protocols are aiming at achieving energy efficiency, throughputs and a better lifetime of the networks. However, two important factors were identified as the major challenges in the Distributed Energy Efficient Clustering (DEEC) protocol. These factors were the distance of the individual nodes and average distance of all from the Base station which were not considered in choosing the cluster heads of the scheme. As a result, the throughputs and the lifetime of the scheme were affected. In this paper, a modified heterogeneous routing protocol called Distance-DEEC (D-DEEC) is proposed. The new algorithm took into account the residual energy, distance of the individual nodes and average distance of all the nodes from the Base station in selecting the cluster heads. This has allowed the protocol to select cluster heads that have high residual energy, closer to the Base station and at the same time not too far from their neighbours. The performance of the proposed algorithm was evaluated using MatLab 2017a and the outcomes showed that D-DEEC protocol outperformed DEEC in terms of energy consumption, throughputs and the network lifetime.

Keywords: Energy Consumption, D-DEEC, Distance, Heterogeneous routing.

1.0 INTRODUCTION

Wireless sensor network (WSN) consists of spatially distributed tiny devices called sensor nodes which cooperatively monitor physical environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants, at different locations (Sharma & Rani, 2014). These nodes collect data about the environment where they are deployed and convey the report directly or indirectly to an external node called Base station (BS) for further examination (kumar & Sudan, 2014).

Heterogeneous routing protocols have been proven to performing better than homogeneous routing schemes (Jibreel et al., 2018); Jibreel, 2019). In these protocols, higher-energy nodes (advanced nodes) which are mostly the cluster heads (CHs) are used to aggregate and transmit the information to the BS while low-energy nodes (normal nodes) are used for sensing and collection of data from the environment. The collected data is relayed to the higher-energy nodes for onward submission to the BS. Some of the protocols proposed in these networks paid much attention to the residual energy of the nodes in selecting the heads but failed to consider their distance to the Base station. A node having higher residual energy helps a lot in data transmission but if its distance from the Base station (BS) is far, then its outcomes may not be appreciative. It is against this background that, this research paper presents an enhanced form of DEEC protocol. The new protocol modified the election probabilities of both advanced and normal nodes by taking into account the distance of the individual nodes and average distance of all the nodes from the Base station in choosing a cluster head. This has helped in improving the throughputs and lifetime of the network.

The remainder of this work is organized as follows: In Section 2, Literature Review; while Methods/Methodology is explained in Section 3; Result is discussed in Section 4 and conclusion is drawn in Section 5**.**

2.0 LITERATURE REVIEW

In literature, several heterogeneous routing protocols have been proposed for WSNs. A few of them are explained below:

Smaragdakis et al. (2004) introduced Stable Election Protocol (SEP). SEP uses two types of nodes namely, normal and advanced nodes. The advanced nodes have more energy than normal ones and therefore advanced nodes are being selected as the cluster heads. This helps in prolonging the stability period and the lifetime of the network. However, the protocol failed to consider the distance of nodes to the BS in picking the cluster heads.

Elbhiri et al. (2010) also presented Developed Distributed Energy-Efficient Clustering (DDEEC) for heterogeneous wireless sensor networks. DDEEC is an advanced version of DEEC. The scheme has been able to resolve penalizing effect in DEEC which they have also noticed but failed to incorporate distance factor in their scheme. The work is similar to Saini et al. (2010) which described Enhanced Distributed Energy Efficient Clustering Scheme (EDEEC), for heterogeneous WSN. EDEEC is another extended version of DEEC with normal, advance and supernode classification based on the node's residual energy. However, a similar challenge was identified as in DEEC algorithm. Saini et al. (2010) further proposed Threshold Distributed Energy Efficient Clustering (TDEEC) protocol. The algorithm is also an improved version of DEEC. TDEEC employed three levels of nodes with a modified probability function. However, the gap between nodes was not taking into account in coming out with the heads. Javaid et al. (2013) proposed Enhanced Developed Distributed Energy-efficient Clustering (EDDEEC) for Wireless Sensor Networks. The protocol is a merger of E-DEEC and DDEEC protocols. The scheme also failed to address distance between nodes issues. This challenge is similar to Jibreel (2018) which explained the Improved Developed Distributed Energy-Efficient Clustering (iDDEEC) scheme. The new algorithm modified the average probability of advanced nodes whose residual energy is less than the threshold residual energy value, **Threv.** iDDEEC implemented TEEN protocol and different amplifica-tion energy levels to conserve energy in the network. The simulation results showed that the proposed protocol performed better than DDEEC in terms of throughputs, residual energy and network lifetime. However, no distance between the nodes and Base station was considered.

Jibreel et al. (2018) proposed a heterogeneous form of Modified Low Energy Adaptive Clustering Hierarchy, Servant-MODLEACH (S-MODLEACH). The algorithm uses three levels of nodes namely, advanced, servant and normal nodes. The protocol chooses Cluster heads based on their residual energy and assigned data aggregation role to a group of nodes called servant nodes. Simulation results showed that S-MODLEACH achieved better outcomes than MODLEACH in respect of throughputs and the network lifetime. The challenge in DDEEC in terms of distance has been identified in this scheme.

An enhanced form of Threshold Distributed Energy Efficient Clustering protocol (TDEEC) has been proposed by Jibreel (2019). The new algorithm, Gateway based-TDEEC, introduced a gateway node at the middle of the sensing area and then installed the BS far away from the sensing field. The cluster heads relay their data to the gateway for aggregation which will then send the final report to the BS. The simulation results showed that the proposed protocol performed better than the TDEEC in terms of stability period, throughput, residual energy and the network lifetime. It however, failed to consider the distance factor is choosing the heads. Therefore, from the literature reviewed, it is clear that, taking into account, the distance of the nodes from the Base station in choosing the cluster heads has not been given the needed attention.

3.0 METHODS/METHODOLOGY

3.1 Existing Protocol

Qing et al. (2006) proposed Distributed Energy Efficient Clustering Protocol (DEEC) protocol. The scheme is a cluster-based method for both multi-level and 2 level energy heterogeneous wireless sensor networks. In the two (2) level energy heterogeneous network, two kinds of nodes are used: normal nodes and advanced nodes. The cluster-heads are elected by a probability based on the ratio between residual energy of each node and the average energy of the network. The nodes with high initial and residual energy (mostly advance nodes) have a better chance of being selected as cluster-heads than the low-energy nodes. This makes DEEC efficient in prolonging the network lifetime, especially the stability period in heterogeneous settings. In DEEC protocol, the average probability (P_i), for both the normal node and the advanced node is given by Equations (1) and (2) respectively.

$$P_i = \frac{P_{opt} E_i(r)}{(1+am)\bar{E}(r)}; \qquad (1)$$

$$\text{if } s_i \text{ is the normal node}$$

$$P_i = (P\_opt(1+a)E\_i(r))/((1+am)E^{-}(r))); \qquad (2)$$

$$(\text{if } s_i \text{ is the advanced node}$$

Where P_{opt} is reference value of average probability p_i ,a , is a constant and m is the percentage of nodes which are considered as advanced nodes.

In homogenous networks, all sensor nodes have the same initial energy so they use P_{opt} to be the reference energy for probability, p_i. However, in heterogeneous networks, the value of P_{opt} is different and it is based on the initial energy of the sensor node.

Also, the average energy $\bar{E}$ (r) of the network for any round r is given by Equation (3)

$$\bar{E}\ (r) = 1/N\ E\_total\ (1 - r/R\ ) \qquad (3)$$

R denotes the total rounds of network lifetime and is calculated using Equation (4)

$$R = \frac{E_{total}}{E_{round}} \qquad (4)$$

$E\_((total))$ is the total energy of the network where $E_{(round)}$ is energy dissipated during each round.

Assuming p_i is the average probability of each sensor node s_i , to become cluster head in a round. During each round, each sensor node chooses a random number between 0 and 1. If the number is less than the threshold as defined in Equation (5) below, the node is eligible to become a CH else not.

$$[T(s]\_(i)\ ) = \begin{cases} \dfrac{p_i}{1 - p_i\left[r \bmod \dfrac{1}{p_i}\right]} & \text{if } s_i \in G \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

Where G is the set of sensor nodes eligible to become CH at round r .

3.2 Drawback in DEEC

These sensor nodes are randomly distributed in the deployment areas. Some of these nodes may be closer or very far from the Base station (BS). The DEEC protocol only considered the residual energy of the nodes in selecting the heads. So when DEEC scheme selects a cluster head and it is far from the Sink, firstly, the head dissipates huge amount of energy in conveying its report to the BS. Secondly, because of the head's signal strength, some of the nodes are compelled to send their data to the distant head for onward transmission to the BS. These nodes also deplete so much energy. As a result, the throughputs and the lifetime of the network in DEEC algorithm are affected. Figure 1 below illustrates this point.

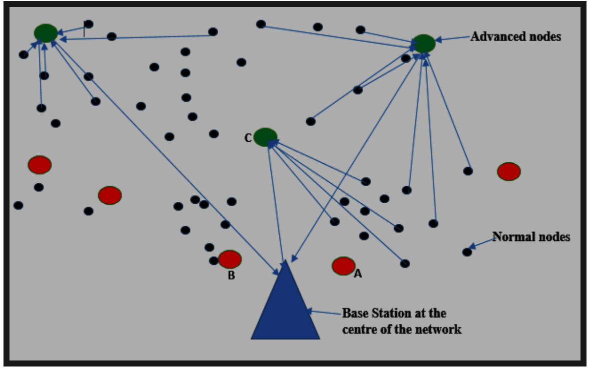


Figure 1: DEEC protocol

Figure1 shows a network where only the residual energy of the node is considered. For instance, when the advanced nodes with the green colours are chosen as cluster heads, the distant nodes are compelled to deplete huge energy to transmit their data to the distant cluster head because of its high received signal strength. The chosen heads, due to the long transmission distance, also waste a lot of energy to convey the final report to the distant Base station. As a result, the throughputs and lifetime of the network are affected.

3.3 Proposed Protocol

The proposed protocol, Distance-DEEC (D-DEEC), implements a similar concept as in DEEC in terms of cluster formation. However, the election probabilities of the protocol have been modified. D-DEEC introduced the ratio of two important factors which are the distance between each node$[\![d]\!]_((il)$ and the BS to the total average distance between all the nodes and BS. In this case, the advanced nodes which have high residual energy, and closer to both the Base Station and their neighbours will rather have a better chance of becoming cluster heads than those which are far from the destination node. This has reduced the energy depletion, the death rate of the nodes and extend the lifetime of the network as a whole. Figure 2 illustrates this point.

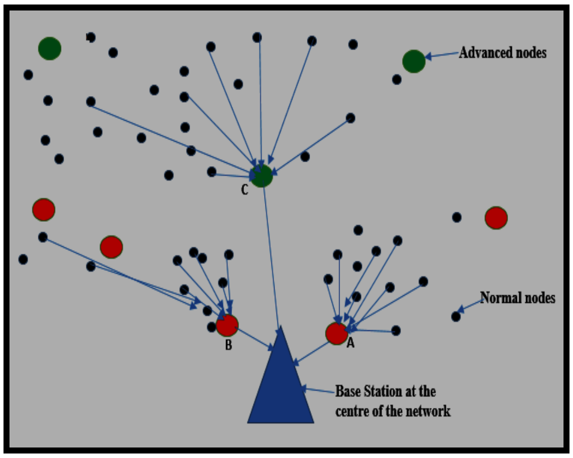


Figure 2: Proposed Scheme (D-DEEC Protocol)

In figure 2, A, B, C are nodes that are chosen as the cluster heads because, they have high residual energy, closer to the Base Station and at the same time not far from their neighbours. This will continue until the chosen cluster heads' energy is reduced to a certain minimum before distant nodes can be considered. With this approach, energy is conserved by the heads and the nodes and throughput is also enhanced.

The new election probabilities of the normal and advanced nodes are given by Equations (6) and (7) respectively.

$$P_i = (P\_opt\ E\_i\ (r))/(((1+am)"E"^{\neg}\ (r))")\ * \frac{d_i}{D_{avg}} \quad \textbf{(6)}$$

$$P_i = (P\_opt\ E\_i\ (r))/(((1+am)"E"^{\neg}\ (r))")\ * \frac{d_i}{D_{avg}} \quad \textbf{(7)}$$

Where d_i is the distance of individual nodes from the BS and D_{avg} is the average distance of all nodes from the BS which is obtained from [2] and it is given by using Equation (8)

$$D_{avg} = \frac{1}{n}\sum_{i}^{n} d_i \quad \textbf{(8)}$$

Each non-cluster-coordinator used energy in relaying k bits data to the cluster head (CH) and is given by Equation (9).

$$E_{non-CHd} = E_{TX}(k, d_{to\ CH}) \quad (9)$$

The total energy spent by each cluster-manager is given by Equation (10)

$$E_{CHd} = \left(\frac{n}{c} - 1\right) kE_{elect} + E_{TX}(k, d_{to\ BS}) \quad (10)$$

Where $d_{to\ BS}$ is the interval from the CH to the BS. The energy used in a group per round is given by Equation (11).

$$E_{cluster\ d} \approx \left(\frac{n}{c} - 1\right) E_{non-CHd} + E_{CHd} \quad (11)$$

The total energy spent by the system is given by Equation (12).

$$E_{total} = c\, E_{cluster\ d}$$

$$E_{total} = c(E\_TX(k,d\_to\,CH) + [kE]\_elect(n/c - 1) + k[n/c\,E]\_DA \;\; + E\_TX(k,d\_to\,BS)) \quad (12)$$

Therefore, the proposed scheme that gives less energy dissipation in transmitting data to the Base station in D-DEEC is given by the Equations (6),(7) and (12).

4.0 RESULTS AND DISCUSSION

In this section, the performance of DEEC and proposed routing protocol, D-DEEC for heterogeneous wireless sensor network have been evaluated using MATLAB 2017a. The experiment performed consisting of 100 nodes randomly deployed in a field of dimension 100m x 100m and a BS located at the centre. All nodes are stationary after deployment and energy loss due to signal collision and interference between signals of different nodes ignored. Table 1 defines the Simulation Parameters used in this research work.

Table 1: Simulation parameters

parameters	values
Network field	(100,100)
Number of nodes	100
Initial energy of normal nodes (E_0)	0.5J
Message size	4000 bits
E_{elec}	50Nj/bit
E_{fs}	10Nj/bit/m^2
E_{mp}	0.0013Pj/ bit/m^2
P_{opt}	0.1

Network Performance parameters taken for analyses are as follows:

i. Network Lifetime
ii. Stability Period
iii. Packets to BS
iv. Packets received by the BS

Figure 1 shows the number of alive nodes during simulation per rounds in DEEC and D-DEEC routing protocols. From the graph, it was observed that the network lifetime was enhanced significantly in D-DEEC compare to DEEC. Nodes were alive up to 3500 rounds in DEEC whilst remain alive up to 5500 rounds in D-DEEC. This means that in D-DEEC scheme, more alive nodes stayed longer than in DEEC routing protocol and thus, making the proposed protocol to have better lifetime than the existing protocol. This is as a result of the distance factor that was taking into account in selecting the cluster heads. The nodes with greater residual energy and closer to their neighbours as well as to the Sink conserved energy better than the far nodes with their neighbours. This has improved the lifetime of the network.

Figure 2 also shows the number of dead nodes per rounds in DEEC and D-DEEC routing protocols. It was noticed from Figure 2 that, the proposed routing protocol has slight better stability period than DEEC protocol. That is, the first node dies at 1000 rounds in DEEC whilst in D-DEEC scheme, it vanished at 1200 rounds. Also, all the nodes are dead at 3500 rounds in DEEC while in D-DEEC, all the nodes died out at 5500 rounds as shown in Figure 2. It indicates that the new scheme has effectively minimized the death rate of nodes. This resulted in lengthier lifespan of the system noticed in the D-DEEC routing scheme.

Figure 3 shows the number of packages transmitted to the Base Station in both the proposed and existing routing protocols. It was realised that the amount of data transmitted to the BS by the DEEC protocol rises from 0 to 3000 rounds and then remains stable throughout the simulation period. Thus, transmitting less amount of data to the BS as shown in Figure 3. In the case of the new algorithm, a large quantity of report was forwarded to the BS from 0 to 4200 rounds and then remain stable. The better performance showed in the new routing protocol is as a result of energy conservation approach that was used and also due to the energetic cluster heads that were selected to transmit data to the Base Station.

Figure 4 shows the number of packets received by the Base Station in both the DEEC and D-DEEC protocols. Sometimes, more data may be transmitted to the BS but a small amount of it may be received by the destination node. This may be due to interference or collision or less energy of the signals during the transmission. It was observed that less number of packets has been received in DEEC protocol whilst the case of the new protocol, a large amount of data was received as seen in Figure 4. This again showed a better performance of the new algorithm in terms of quantity of report received by the Base station.

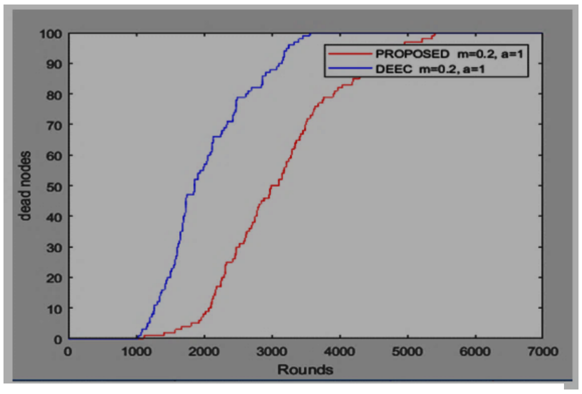


Figure 1: Number of Alive Nodes per round

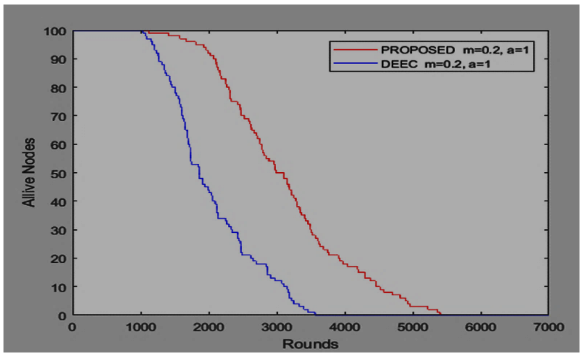


Figure 2: Number of Dead Nodes per round

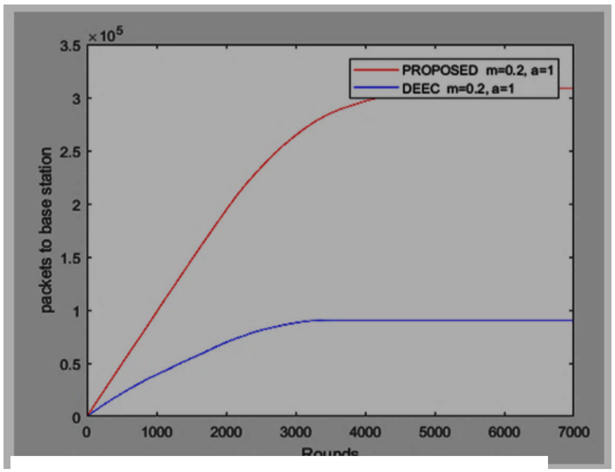


Figure 3: Packets to the Base Station

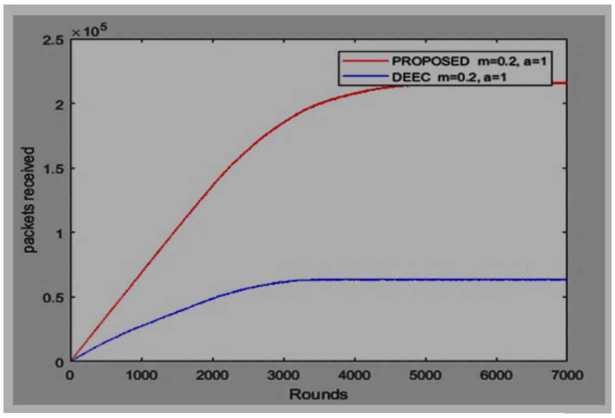


Fig 4: Packets received by the Base Station

5.0 CONCLUSION

A new heterogeneous routing protocol called Distance-DEEC (D-DEEC) has been proposed in the research work. The proposed scheme modified DEEC protocol by introducing the distance of the individual nodes and average distance of all the nodes from the Base Station in its election probabilities. The approach has minimized the energy depletion of the nodes and made them stayed alive longer and transmit more data to the Base station than the existing scheme. The simulation was conducted to evaluate the performance of the new scheme and the simulation results showed that the D-DEEC protocol outperformed DEEC protocol in terms of stability period, throughputs and lifetime of the network.

REFERENCE

Elbhiri, B., Saadane, R., El Fkihiand S., & Aboutajdine, D. (2010). Developed Distributed Energy-Efficient Clustering (DDEEC) for heterogeneous wireless sensor networks", I/V Communications and Mobile Network (ISVC), 5th International Symposium on, pp.1-4

Javaid, N., Qureshi, T.N., Khan, A.H., Iqbal, A., Akhtar, E., & Ishfaq, M. (2013). EDDEEC: Enhanced Developed Distributed Energy-efficient Clustering for Wireless Sensor Networks, Procedia Computer Science, Volume 19, pp. 914-919.

Jibreel, F. (2018). Improved Developed Distributed Energy-Efficient Clustering Scheme (iDDEEC**).** International Journal of Innovative Science and Research Technology, Volume 3, Issue 12, pp. 564-567.

Jibreel, F. (2019). Gateway–based Threshold Distributed Energy Efficient Clustering (G-TDEEC). *International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 42, pp. 43-46.*

Jibreel, F., Daabo, M.I., Yusuf-Asaju, A.W., & Gbolagade, K.A. (2018). Servant-MODLEACH Energy Efficient Cluster Based Routing Protocol for Large Scale Wireless Sensor Network. *The 12th International Multi-Conference on ICT Applications*, Volume XII, pp. 1-6

kumar, L., & Sudan, M. (2014). Advancement in Single Node Data Transfer Energy Protocol Using Gateways in Wireless Sensor Network. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, pp. 106-113

Qing, L., Zhu, Q., & Wang, M. (2006). Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks, Computer Communications, vol. 29, no.12, pp. 2230-2237.

Saini, P. & Sharma, A. K. (2010). Energy Efficient Scheme for Clustering Protocol Prolonging the Lifetime of Heterogeneous Wireless Sensor Networks. International Journal of Computer Applications 6(2), pp. 30-36.

Saini, P., & Sharma, A. K. (2010). E-DEEC-Enhanced Distributed Energy Efficient Clustering Scheme for heterogeneous WSN. 1st International Conference on Parallel, Distributed and Grid Computing, pp. 205-210.

Sharma, S., & Rani, M. (2014). A Survey and Comparative Study of Routing Protocols in Wireless Sensor Network. International Journal of Engineering and Computer Science, Volume 3 Issue 5, pp. 6008-6012

Smaragdakis, G., Matta, I., & Bestavros, A. (2004). SEP: A stable election protocol for clustered heterogeneous wireless sensor networks, Proc of the Int'1 Workshop on SANPA, pp. 251-261.

CUSTOMERS CHURN PREDICTION IN FINANCIAL INSTITUTION USING ARTIFICIAL NEURAL NETWORK

[1*]Amuda K.A. and [2]Adeyemo A. B.
Department of Computer Science, University of Ibadan, Nigeria.
*Email of Corresponding Author: *akindeleamuda@gmail.com*

ABSTRACT

In this study a predictive model using Multi-layer Perceptron of Artificial Neural Network architecture was developed to predict customer churn in financial institution. Previous researches have used supervised machine learning classifiers such as Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Random Forest. This study developed a model that eliminate the use of feature selection techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Correlation matrix with heatmap and feature importance in data preprocessing stage. Fifty thousand customers' data were extracted from the database of one of the leading financial institution in Nigeria for the study. The multi-layer perceptron model was built with python programming language and used two overfitting techniques (Dropout and L2 regularization).

The implementation was carried out using python and Neuro solution infinity software. The results showed that the Artificial Neural Network software developed (in Python) had comparable performance with that obtained from the Neuro Solution Infinity software. The accuracy rates are 97.53% and 97.4% while ROC (Receiver Operating Characteristic) curve graphs are 0.89 and 0.85 respectively.

Keywords: Multi-layer Perceptron, Artificial Neural Network, Customer Churns, Classifiers, L2 regularization and Dropout.

1.0 INTRODUCTION

The volume of data has been growing at a fast rate due to technological advancement which makes data as the oil of the 21st century but oil is just useless until it refines into fuel. Many new methods and techniques have been introduced to extract the meaningful and salient information hidden in the data and the process is called Data Mining Techniques. The technique included Support Vector Machine (SVM), Linear Regression, Genetic algorithm, Decision Tree and Neural Network (Kolajo & Adeyemo 2012).

Customer churn is a fundamental problem for companies and it is defined as the loss of customers because they move out to competitors. Being able to predict customer churning behavior in advance, gives an institution a high valuable insight in order to retain and increase their customer base.

A wide range of customer churn predictive models have been developed in the last decades. Most advanced models make use of state-of-the-art machine learning classifiers such as random forest, linear and logistic regression (Castanedo et al, 2014).

One of the most direct and effective approaches to keep the current customers is that the company should be able to foresee potential churners in time and react to it quickly. Recognizing the indications of potential churn; satisfying customer needs, restoring and re-establishing loyalty are actions supposed to help the s organization minimize the cost of gaining new customers (Mitkees et al, 2017).

Financial institutions are the organization that process monetary transaction including business and private loans, customer deposits and investments. As customers are directly related to profits, financial institutions must avoid the loss of customers while acquiring new customers. Harvard Business Review believes that by reducing the customer defection rate by 5%, companies can increase profits by 25% to 85%, while Business Week thought the profits will increase by 140%.

Bhattacharya (1994) that the cost of developing a new customer is 5 to 6 times than retaining an old customer. As can be seen, reducing customer attrition has a significant impact not only on increasing profits for financial institutions, but also on enhancing their core competitiveness. Therefore, it is urgent for financial institutions to improve the capabilities to predict customer churn, thereby taking timely measures to retain customers and preventing other clients from churning.

Artificial Neural Network also called Neural Network (NN) is a complex network that comprises a large set of simple nodes known as neural cells. Artificial Neural Network was proposed based on advanced biology research concerning human brain tissue and neural system, and can be used to simulate neural activities of information processing in the human brain (Vaxevanidis,2008).

Artificial Neural Network have been used in finance for portfolio management, credit rating and predicting bankruptcy, forecasting exchange rates, predicting stock values, inflation and cash forecasting.

In this paper, we work on predictive analysis of churning behavior in financial institution base on the dataset extracted from the bank database. The model developed with MLP of Artificial Neural Network architecture, potential churners and non-churners were predicts and different performance metrics such as precision, accuracy, recall and f-score were used to evaluate the robustness of the model and compared with Neuro Solution Infinity analytical tool. The paper is structured as follows. Section 2 explores the various literature reviews on the related work. Section 3 gives the research objectives and illustrates the methodological approach. Section 4 illustrates the results of the proposed work and Section 5 explores the conclusion and future work.

2.0 LITERATURE REVIEWS

This section investigates on the related work which has been on customers churn prediction in the different domain likes telecommunication and banking industry using different supervised machine learning techniques.

He et al, (2014) carried out a research on customer attrition analysis of commercial bank using Support Vector Machine. The dataset contains 50000 customers records were extracted from Chinese commercial bank data warehouse. After the removal of missing values and outliers, 46,406 records used to model. SVM algorithm was applied but due to the imbalanced characteristics of the dataset, Random sampling method was introduced to improve SVM since it has a higher degree of recognition. The results shown that the combination of random sampling and support vector machine algorithm significantly improved the predictive power and accurately predict churning rate.

In the telecommunication industry, cellular network providers are becoming more competitive and churn management has become a crucial task telecommunication industry. Sharma, (2011) applied a neural network to predict customer churn in cellular network service. The dataset contains 20 variables worth of information about 2,427 customers were collected from UCI Machine Learning Database at the University of California, Irvine. Neural Network was implemented on Clementine data mining software package from SPSS Inc., Clementine provided two different classes of supervised neural networks namely Multilayer Perceptron (MLP) and Radial Base Function Network (RBFN). The result shows that the model predicts customers churn with an accuracy rate of more than 92%.

Oyeniyi & Adeyemo, (2015) customer churn has become a major problem within a customer-centered banking industry and banks have tried to track Customer Interaction in order to detect early warning signs in customer's behaviour such as a reduction in transaction and account dormancy. They also worked on customer churn analysis in the banking sector, the model used K-Means and Repeated Incremental Pruning to Produce Error Reduction (JRip algorithm) which was implemented on Weka. The dataset was extracted from the bank's customer relationship management database and transaction warehouse from a major Nigeria bank. The results determine the pattern in customer behavior and help banks to identify customers that are likely to churn.

To improve the prediction abilities of machine learning method, Guo-en et al (2008) applied support vector machine on structure risk minimization to customer churn analysis. The dataset was collected from machine learning UCI database of the University of California. The results shown that the support vector machine method enjoys the best accuracy rate, hit rate, covering rate and lift coefficient when compared with Artificial Neural Network, Decision Tree, Logistic Regression and Naïve Bayesian classifier. Support vector machine provides an effective measurement for customer churn prediction.

Kolajo & Adeyemo (2012), used descriptive and predictive data mining techniques to determine the calling behaviour of subscribers and recognize subscribers with a high probability of churn in a telecommunication company. In the descriptive stage, the customers were clustered based on their usage behavioural features and algorithms used for clustering methods were K-Means and Expected Maximization (EM). In the predictive stage, DecisionStump, M5P and RepTree classifiers algorithms were implemented in Weka. The results show that EM performs better than K-mean in the descriptive stage while M5P perform better than both DecisionStump, and RepTree in the predictive stage.

Lu et al, (2011) proposed the use of boosting algorithm to predict the customer churn in the telecommunication industry. Customers were separated into two clusters based on the weight assigned by boosting algorithm, the data set used was extracted from a telecommunication company which included a segment of mobile customers who are active and contains about 700 variables. Boosting algorithm performs better than logistic regression and it provides a good separation of churn data.

Sandeepkumar & Monica (2019), considered multi-layered neural network which also known as Deep Feed Forward Neural Network (DFNN) to perform predictive analytics on customer attrition in the banking sector. The dataset was collected from the UCI machine learning archive which has total 10,000 customer data with 14 dimensions of features. The model use optimized one hot encoding and Tukey outliers' algorithms to perform data cleaning and preprocessing. The model was compared with the machine learning algorithms such as Logistic regression, Decision tree and Gaussian Naïve Bayes algorithm. The results shown that Enhanced Deep Feed Forward Neural Network (DFNN) model performs best in accuracy when compared with others machine learning algorithms.

Wang et al (2018), worked on large-scale ensemble model for customer churn prediction in search Ads. The aimed of the research was to detect customers with a high propensity to leave the ads platform. the ensemble model of gradient boosting decision tree (GBDT) was used to predict customer that will be a churner in the foreseeable future based on its activities in the search ads. two different features for the GBDT were dynamic features and static features. dynamic features considered a sequence of customers' activities such as impressions, clicks during a long period. while static features considered the information of customers setting such as creation time, customer type. the dataset was collected from Bing Ads platform and the result shown that the static and dynamic features are complementary with AUC (area under the curve of ROC) value 0.8410.

Rosa (2019), proposed a new framework that used Artificial Neural Network for assessing and predicting customer attrition in the banking industry. About 1588 customers data from the period of January 2017 to

December 2017 was extracted from the bank´s Data Warehouse with the use of SAS Base. The research solely focused on developing neural network and overlooked other machine learning algorithms such as Decision Trees, Logistic Regression, or Support Vector Machines.

3.0 MATERIALS AND METHODS

Customer churn prediction has a great impact on financial institutions since they are mainly depending on customer satisfaction for their operations. These institutions are highly competitive environment and retain customers by satisfying their needs under the resource's constraint. The data mining techniques used to discover interesting patterns and relationships that exist in the data and predict the customers' behavior either churning or non-churning by fitting the model on the available historical data.

Researches on churn prediction have used machine learning classifiers such as Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Random Forest. These classifiers use techniques such as Recursive Elimination Feature (RFE), Principal Component Analysis (PCA), Correlation matrix with heatmap and feature importance which lead to over-specified and incomplete feature selection. This study developed a model that eliminate the use of feature selection techniques in data preprocessing stage.

The aim of this study is to develop customer churn predictive model using Multilayer Perceptron of Artificial Neural Network Architecture. The model designed to predict the potential churners and non-churners and compare the performance with proprietary software.

3.1 Data Collection

The dataset was extracted from the database of one of the leading financial institution in Nigeria. The contained about 50000 customers' data with 42 attributes. Some of these are:

Features	**Description**
Gender	M for Male; F for Female and NULL was used for customers we couldn't get their gender type
Cust_txn_status	Active and Inactive
Marital_status	D= Divorced; M= Married; S=Single, NULL=NULL
Occupation	Customer Job description
Lga	Local Government Area
State	State in Nigeria
Religion	Christian; Islam; Other Religion
current_account	Current Accounts for individuals
xclusive_subscript	Current Account for High Net worth Individuals with monthly income above N1m
current_account_corp	Current account for big companies like shell, chevron, coca-cola, nestle.
savings_deposit_youth	Savings accounts for teenagers and babies
Community_savings_account	Savings account for estates and communities
Hida	High interest deposit account
Mobapp_Fund_Trsf_LCY_Revenue	Profit accruable to the Bank based on currency transfers by the customer between Sept to November
Mobapp_Fund_Trsf_FCY_Count	The total number of times the customer transferred money (USD or EURO or GBP only).
Mobapp_Fund_Trsf_FCY_Vol	The total amount of money the customer transferred (USD or EURO or GBP only)
Mobapp_Fund_Trsf_FCY_Revenue	Profit accruable to the Bank based on foreign currency transfers made by the customer between Sept to November
Mobapp_Lifestyle_Count	The total number of times the customer paid for travel tickets or movie tickets or other event ticket
Mobapp_Lifestyle_Vol	Amount in naira times paid by customer for travel tickets or movie tickets or other event ticket.
mobapp_Lifestyle_Revenue	Profit accruable to the bank based on transactions made by customers for travel tickets or movie tickets or other event.
Cust_id	Unique identification number for customers.

3.2 Data Preparation

The missing values with 30% null were removed from the dataset with the aid of Python programming language libraries. Numerical data was replaced with the 'mean' of the variables while the 'mode' was used for the categorical data. To achieve better performance, the categorical data was transformed to numerical format using the Label Encoder function in Python. Feature scaling was applied to normalize the data and improved the computational time.

3.3 Tools and Libraries

Anaconda: is an open source software distribution of R and Python programming languages that are used for scientific computing such as data science, predictive analytics, machine learning, and deep learning applications purposely to simplify package management and deployment.

Jupyter Notebook: is an open free source web application that is used for data cleaning and transformation, numerical, simulation, statistical modelling, data visualization and so on.

Matplotlib: is an amazing visualization library in Python programming language for two-dimensional plots of arrays. One of the greatest benefits of visualization is that it allows high dimensionality data to visualize and easily understandable and it consists of several plots like line, bar, scatter, histogram and so on.

Pandas: is the most popular Python programming language package that offers powerful, expensive and flexible data structures that make data manipulation and analysis easy.

Numpy: is the fundamental package for scientific computing in Python programming language that contains a powerful N-dimensional array object and also useful in linear algebra.

Label Encoder: is a Python programming language package that is used to transform non-numerical labels (or nominal categorical variables) to numerical labels.

ANN visualizer: is a Python programming language library that enables visualization of an artificial neural network.

Seaborn: is a Python data visualization library based on matplotlib that provides high-level interface for drawing attractive and informative statistical graphics. **Sci-kit Learn**: it is a free machine learning library for Python programming language that designed to interoperate with Python numerical Numpy and scientific libraries SciPy. Also, it can be used for classification, regression and clustering algorithms including support vector machine, linear and logistic regression, random forests, gradient boosting, decision tree, K-means and so on.

TensorFlow: is an open source Python library used for machine learning applications such as neural network and used Keras as a backend.

Keras: is a neural network framework for Python programming language that provides a convenient way to define and train almost any kind of deep learning model.

Neuro Solutions Infinity: is one of the most powerful neural network software of the Neuro Solutions family that streamlines the data mining process by automatically cleaning and preprocessing data. It uses distributed computing, advanced neural networks and artificial intelligence (AI) to model data; creates highly accurate predictive models with an easy-to-use and intuitive interface that provides valuable insights that can be used to drive better decisions.

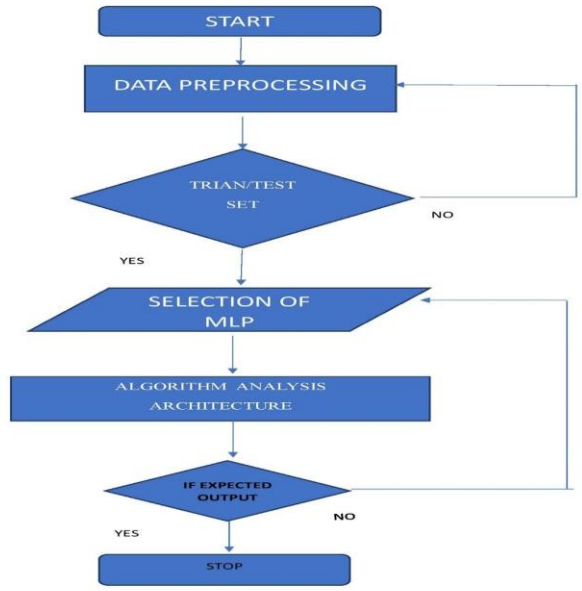


Figure 1: Flow Diagram for proposed model

Microsoft SQL Server: is a relational database management system developed by Microsoft Incorporation which primarily used for storing and retrieving data requested by other software applications that run on the same computer or on another computer across a network.

Flowchart: illustrate the step by step procedure on how to predict customers churn with the multilayer perceptron of artificial neural network architecture.

3.4 Implementation

The implementation was carried out in two phases; MLP software development on Python and then using Neuro Solution Infinity software. Multilayer Perceptron: MLP is an architecture of the artificial neural network that consists of multiple layers where each layer is fully connected with the next layer in a feed-forward direction. The first layer and the last layers represent the inputs (independent variables) and outputs (target variables) of the system respectively. Connections between the nodes represented as weights. The more the numbers of hidden layers in the network, the more the complexity of the network.

In multilayer perceptron architecture, each hidden layer node consists of two parts;

- Summation function: The summation function calculates the sum of each input value multiplied by the corresponding weights. Mathematically, it can be represented as

$$S = \sum_{i=0}^{n} Wi\, Xi \qquad (1)$$

Where n is the numbers of neuron in the network

- Activation function: The activation is applied by each neuron to sum of weight input signals in order to determine its output signal. Sigmoid function such as the S-shaped curve is one of the most commonly used activation functions for binary classification problems.

$$\emptyset(S) = \frac{1}{1+\exp(-S)} \qquad (2)$$

3.5 Steps for MLP Model

(i) Extraction of customer data including transactional history from bank database.

(ii) Data pre-processing for removal of noise and transformation of categorical data into numerical data.

(iii) Data splitting (train and test set) and feature scaling

(iv) Start at the input layer by forwarding propagates patterns of the train data through the network to generate output.

(v) Using cost function to calculate the network output to minimize the error rate.

(vi) Find the derivative with respect to each weight in the network and updating the model,

(vii) Calculate the network output and apply threshold function to obtain the predicted class label.

(viii) Evaluate the model.

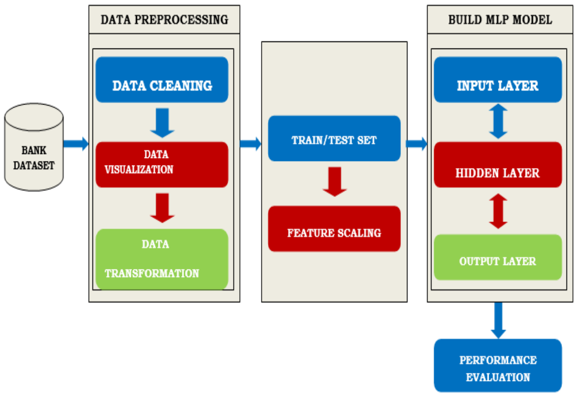
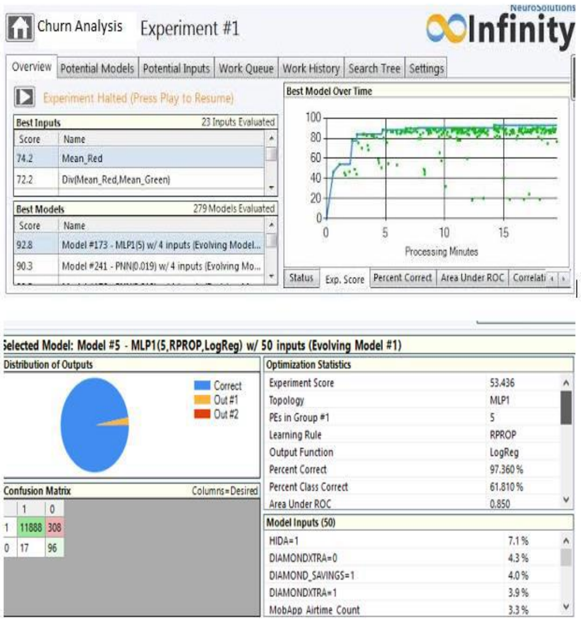


Figure 2: Proposed Model of Multilayer Perception of Artificial Neural Network.

Neuro Solution Infinity Software: The implementation was done by launching the software followed by uploading the data from the Personal Computer (PC) into the software platform. The dataset divided into three parts; training set, test set, and validation set and the threshold for the missing values was set. The Neuro Solution Software comprises of different algorithms such as Multilayer Perceptron (MLP), Probabilistic Neural Network (PNN), and Support Vector Machine (SVM). In figure 3 about 50 different models ran on the software and the best model was selected for the churn analysis.



Best Inputs 23 Inputs Evaluated

Score	Name
74.2	Mean_Red
72.2	Div(Mean_Red,Mean_Green)

Best Models 279 Models Evaluated

Score	Name
92.8	Model #173 - MLP1(5) w/ 4 inputs (Evolving Model...
90.3	Model #241 - PNN(0.019) w/ 4 inputs (Evolving Mo...

Selected Model: Model #5 - MLP1(5,RPROP,LogReg) w/ 50 inputs (Evolving Model #1)

Confusion Matrix Columns=Desired

	1	0
1	11888	308
0	17	96

Optimization Statistics

Experiment Score	53.436
Topology	MLP1
PEs in Group #1	5
Learning Rule	RPROP
Output Function	LogReg
Percent Correct	97.360 %
Percent Class Correct	61.810 %
Area Under ROC	0.850

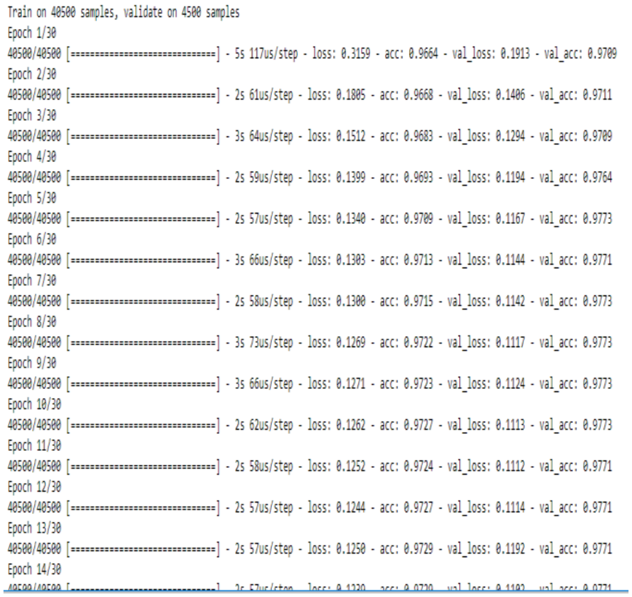
Model Inputs (50)

HIDA=1	7.1 %
DIAMONDXTRA=0	4.3 %
DIAMOND_SAVINGS=1	4.0 %
DIAMONDXTRA=1	3.9 %
MobApp Airtime Count	3.3 %

Figure 3: Screenshot of Neuro Solution Infinity Interface.

4.0 RESULTS AND DISCUSION

The result of the implementation in Python was presented in figure 4, figure 5 and figure 6. It can be seen that training loss was decreasing while training accuracy was increasing and also the validation loss was decreasing while validation accuracy was increasing. These shown that the model was neither overfitting nor underfitting.



```
Train on 40500 samples, validate on 4500 samples
Epoch 1/30
40500/40500 [==============================] - 5s 117us/step - loss: 0.3159 - acc: 0.9664 - val_loss: 0.1913 - val_acc: 0.9709
Epoch 2/30
40500/40500 [==============================] - 2s 61us/step - loss: 0.1805 - acc: 0.9668 - val_loss: 0.1406 - val_acc: 0.9711
Epoch 3/30
40500/40500 [==============================] - 3s 64us/step - loss: 0.1512 - acc: 0.9683 - val_loss: 0.1294 - val_acc: 0.9709
Epoch 4/30
40500/40500 [==============================] - 2s 59us/step - loss: 0.1399 - acc: 0.9693 - val_loss: 0.1194 - val_acc: 0.9764
Epoch 5/30
40500/40500 [==============================] - 2s 57us/step - loss: 0.1340 - acc: 0.9709 - val_loss: 0.1167 - val_acc: 0.9773
Epoch 6/30
40500/40500 [==============================] - 3s 66us/step - loss: 0.1303 - acc: 0.9713 - val_loss: 0.1144 - val_acc: 0.9771
Epoch 7/30
40500/40500 [==============================] - 2s 58us/step - loss: 0.1300 - acc: 0.9715 - val_loss: 0.1142 - val_acc: 0.9773
Epoch 8/30
40500/40500 [==============================] - 3s 73us/step - loss: 0.1269 - acc: 0.9722 - val_loss: 0.1117 - val_acc: 0.9773
Epoch 9/30
40500/40500 [==============================] - 3s 66us/step - loss: 0.1271 - acc: 0.9723 - val_loss: 0.1124 - val_acc: 0.9773
Epoch 10/30
40500/40500 [==============================] - 2s 62us/step - loss: 0.1262 - acc: 0.9727 - val_loss: 0.1113 - val_acc: 0.9773
Epoch 11/30
40500/40500 [==============================] - 2s 58us/step - loss: 0.1252 - acc: 0.9724 - val_loss: 0.1112 - val_acc: 0.9771
Epoch 12/30
40500/40500 [==============================] - 2s 57us/step - loss: 0.1244 - acc: 0.9727 - val_loss: 0.1114 - val_acc: 0.9771
Epoch 13/30
40500/40500 [==============================] - 2s 57us/step - loss: 0.1250 - acc: 0.9729 - val_loss: 0.1192 - val_acc: 0.9771
Epoch 14/30
40500/40500 [==============================] - 2s 57us/step - loss: 0.1230 - acc: 0.9729 - val_loss: 0.1192 - val_acc: 0.9771
```

Figure 4: Learning Process

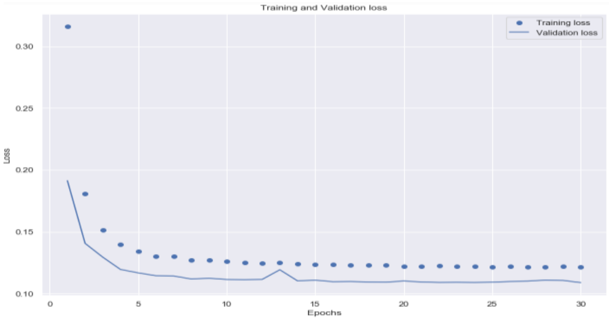


Figure 5: Training Loss and Validation Loss

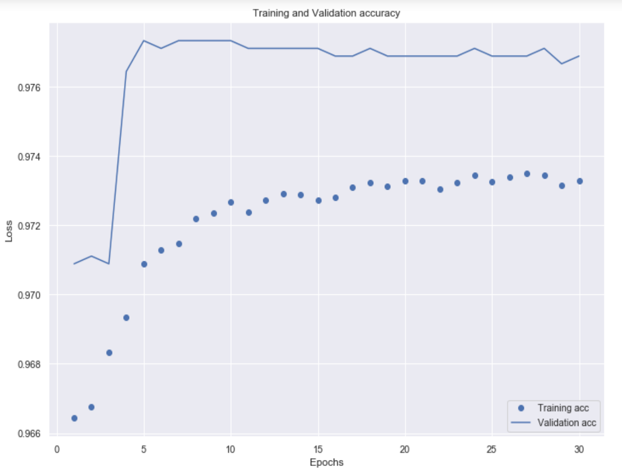


Figure 6: Training Accuracy and Validation Accuracy

Performance Evaluation: Confusion matrix was used to describe the performance of the Multilayer perception classifier on a set of test data for which the true values are known. It aids the visualization of the performance of the model. The confusion matrix table summarized the prediction results of the model. The numbers of correctly classified and incorrectly classified (misclassified) are summarized with count values and broken down by each class (churners and non-churners). The essence of performance evaluation is to gives insight not only into the errors being made by the model but more importantly the types of errors that are being made.

ROC curve was used to know the capability of the models to distinguish between classes (churners and non-churners). ROC values range between 0 and 1, that is, if the value close to 1, it means the model has good measure of separability while near 0 means it has worst measure of separability (i.e. it is reciprocating the result). Figure 7 it can be seen that the model correctly classified 38 customers as churners, 4841 as non-churners and incorrectly misclassified 9 customers as churners and 112 as non-churners.

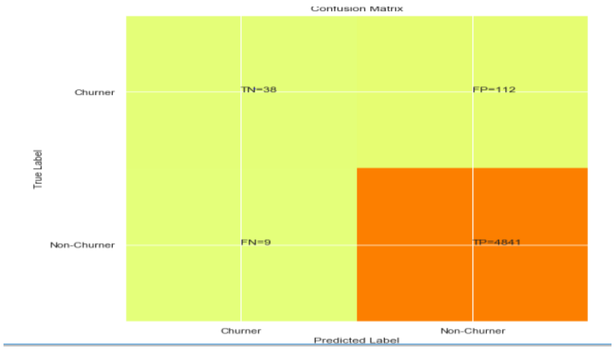


Figure 7 Confusion Matrix on Python

From figure 8 the value for ROC curve graph was 0.89 which mean the model was working perfectly. Dropout was introduced in the network to reduce overfitting and the complexity of the network to the minimal level.

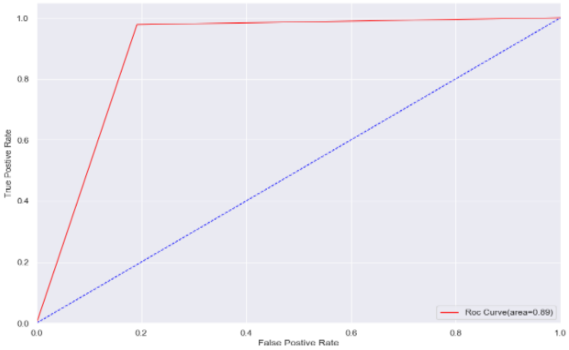


Figure 8: ROC Curve on Python

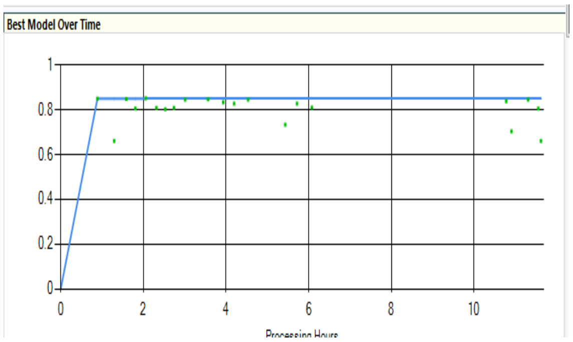


Figure 9: Neuro Solution Infinity ROC Curve

Figure 9 shows the ROC curve for the best model in the Neuro Solution Infinity software, Multilayer Perceptron (MLP) model gave the best result with 0.85 followed by Probabilistic Neural Network (PNN), 0.84 and Support Vector Machine (SVM), 0.842.

4.1 Comparative Analysis of Implementations

Table 1 presents the results of the two implementations; it can be seen that the Multilayer Perceptron implemented on Python's result was comparable to that of the Neuro Solution Infinity software.

	Python	**Neuro Solution Infinity Software**
Accuracy Rate	97.53	97.36
ROC curve Value	0.89	0.85
Precision	97.7	97.6
Recall	99.8	99.9
F – Measure	98.8	98.7

Table 1: Implementation Results

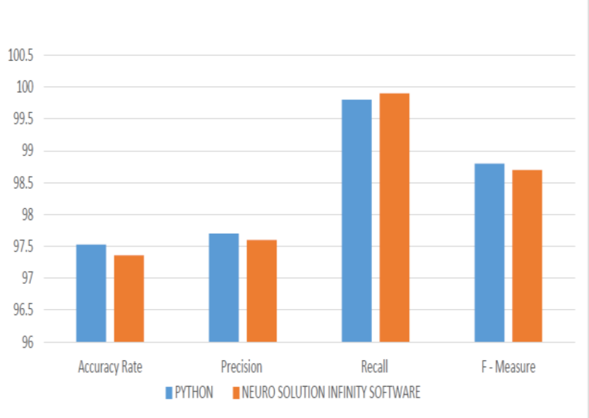


Figure 10: Pictorial view of the two implementations performance.

Dropout was introduced in the network to reduce the degree overfitting and the complexity of the network to the minimal level.

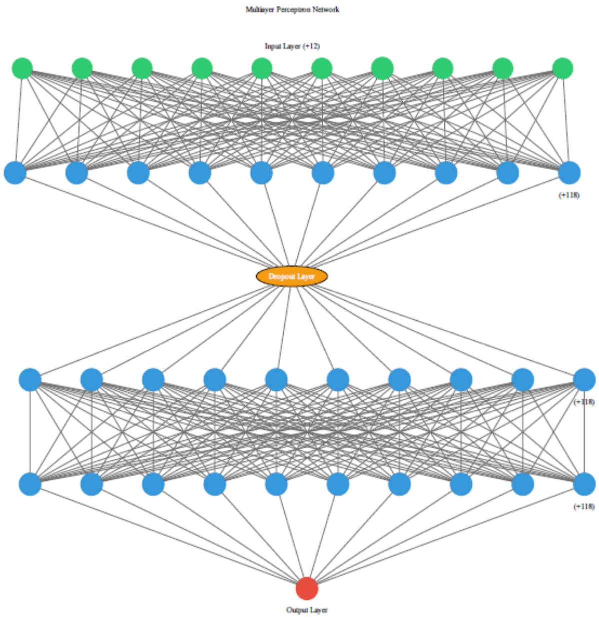


Figure 11: Schematic Diagram of the Proposed Multilayer Perceptron (MLP) model

5.0 CONCLUSION

This study was designed for the prediction of customers churn using the data from a financial institution in Nigeria. The data was extracted from the bank database and divided into three sets: training set, test set, and validation set. Eighty percent (80%) of the dataset was used for training, ten percent (10%) was used for testing and the remaining ten percent (10%) was used for validating the model. Data pre-processing was carried out by removing thirty percent (30%) of missing values and the text (categorical) data were converted into numerical data. Also, feature scaling was employed to increase the computation time of the algorithm.

Multilayer Perceptron Artificial Neural Network architecture was implemented on two different platforms; Python and Neuro Solution Infinity Software. The results obtained using the two software were comparable with the Python implementation given a 97.53 % accuracy while the Neuro Solution Infinity software gave 97.36% accuracy. The performance of the models was evaluated on the test dataset (unseen data) using the following metrics: Accuracy, Precision, Recall, F-measure. In recent surveys Python programming language has been gaining in popularity as the tool of choice amongst Data Scientists and Data Mining and Natural Language researchers. This is due to its being relatively easy to learn and its very rich library of functions and utilities. The information obtained from the predictive model can be used for decision making in customer retention management system. For future work other ANN architectures such as Convolution Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) can be used for customers churn prediction.

REFRENCES

Bhattacharya, C. B. (1994). When Customers Are Members: Customer Retention in Paid Membership Contexts?

Castanedo, F. (2014). Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network,1-8 https://www.semanticscholar.org/paper/Using-Deep-Learning-to-Predict-Customer-Churn-in-a-Castanedo/7e2eba72e678ed796cd4c69b53ed0be4736e8b0c

Guo-en, X. I. A., & Wei-dong, J. I. N. (2008). Model of Customer Churn Prediction on Support Vector Machine. *Systems Engineering - Theory & Practice*, *28*(1),71–77. https://doi.org/10.1016/S1874-8651(09)60003-X

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia - Procedia Computer Science*,*31*,423–430. https://doi.org/10.1016/j.procs.2014.05.286

Lu, N., Lin, H., Lu, J., & Zhang, G. (2011). A Customer Churn Prediction Model in Telecom Industry Using Boosting, (c), 1–7.

Mitkees, I. M. M., Ibrahim, A., & Elseddawy, B. (2017). Customer Churn Prediction Model using Data Mining techniques, 262–268.

Oyeniyi, A. O., & Adeyemo, A. B. (2015). Customer Churn Analysis in Banking Sector Using Data Mining Techniques, *8*(3), 165–174.

Rosa, Nelson Belém da Costa. *Gauging and foreseeing customer churn in the banking industry: a neural network approach*. Diss. 2019

Sandeepkumar hegde, & Mundada, M. R.(2019). *Enhanced Deep Feed Forward Neural Network Model for the Customer Attrition Analysis in Banking Sector*. (July), 10–19. https://doi.org/10.5815/ijisa.2019.07.02

Sharma, A. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services, *27*(11), 26–31.

Kolajo, Adeyemo. (2012). Computing, Information Systems Data Mining Technique for Predicting Telecommunications Industry Customer Churn Using both Descriptive and Predictive Algorithms Customer Churn Using both Descriptive and Predictive Algorithms, *3*(2), 27–34.

Wang, Q., Xu, M., & Hussain, A. (2018). *Large-scale Ensemble Model for Customer Churn Prediction in Search Ads*.

Vaxevanidis, N. M. (2008). Artificial neural network models for the prediction of surface roughness in electrical discharge machining, 283–292. https://doi.org/10.1007/s10845-008-0081-9.

A ONE FACTOR PERFORMANCE ANALYSIS TO PREDICT ACCIDENT CASUALTY USING DATA MINING TECHNIQUES

[1]Abdulfatah S. M. and [2]Babangida A. A.
[1] Department of Computer Software Engineering, Katsina State Institute of Technology and Management, Nigeria
[2]Department of Mathematics and Computer Science, Umaru Musa Yar'adua University, Katsina State, Nigeria.
Email of Corresponding Author: abdulfatahsamaila@gmail.com

ABSTRACT

Road accident is one of the undesirable events that causes unnatural deaths, disability and property damage. Researching and understanding of the causes of road accident to predict its occurrence and put in place necessary preventing measures is vital. One Factor Prediction Algorithm (OFPA) for car accidents were proposed to predict occurrence and casualty types. Classification and prediction techniques of Data Mining were applied to identify hidden patterns in the accident data set to find the determinant factors that contribute to the occurrence of accidents. To achieve these goals, the CRISP-DM 1.0 standard data mining methodology has been adopted and the WEKA data mining tool was used to implement the proposed algorithm using J48 and Sequential Minimal Optimization (SMO) algorithms. The results of the paper have shown that J48 Algorithm performed best in the analyses. The comparison of the model using WEKA's experimenter also showed that J48 outperforms SMO algorithm of One Factor Prediction Algorithm (OFPA).

Keywords: Data Mining, Accident, One Factor Predicting Algorithm, J48 Algorithm, Sequential Minimal Optimization (SMO).

1.0 INTRODUCTION

Han, Jiawei, Kamber, and Micheline. (2006) simply stated, “Data mining refers to extracting or “mining” knowledge from large amounts of data”. There are some other terms which carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data or pattern analysis, and data archaeology. This fast growth and tremendous amount of data, collected and stored in large and numerous databases need a powerful tool to elicit useful information. The tool helps to get benefit from the collected data, by identifying relevant and useful information. Data mining is one of the solutions to analyze huge amount of data and turn such data into useful information and knowledge (Han, Jiawei, Kamber, & Micheline., 2006).

In general data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, Han, Jiawei, Kamber, and Micheline. (2006) classified data mining tasks into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database where as predictive mining tasks perform inference on the current data in order to make predictions.

Road accident is one of the undesirable events that are uncertain but predictable. Road accident is one of the major causes of unnatural deaths, disability and property damage (Sachin & Durga, 2015).

The aim of this paper is to identify Determinant factors that contribute causes of accident and to develop accident prediction model that improves road safety.

2.0 LITERATURE REVIEW

In their research work Gwamki, Liman, and Rumah (2018) designed to examined the role played by the Nigerian Federal Road Safety Corps (FRSC) in the management of Road Traffic Crashes in Katsina metropolis; they identified some of the causes of the crashes in the area like the lack of adequate training of drivers, the use of poor tires by drivers.

Other causes include; driving under the influence of drugs, poor roads as well as poverty and illiteracy on the part of majority of the drivers account for some of the major reasons responsible for the accidents experienced in Katsina metropolis (Gwamki et al., 2018).

Addi, Tarik, and Fatima (2016) used Large Scale Data Mining method especially based on the association rules and multi criteria analysis approach to discover new knowledge from historical data about traffic accidents in one of morocco busiest roads in order to assist police decision makers in the formulation of new policies and traffic rules on our highways management.

Addi, Gharnati, and Tarik (2017) found out in many countries, road transport often involves accidents, and this affects transport and shipping services. Understanding road traffic is extremely important in improving road safety. In their research, they propose an effective method for mining strong and relevant association rules from a road accident database. With the objective of identifying the hidden relationships between the most common accidents, the road accident dataset is analyzed using the association rules technique.

Sachin and Durga (2015) exploit the Association Rule Mining method to analyze accident patterns for different types of accidents on the road. The method was applied on this base on clusters to identify some interesting rules that can help in understanding the circumstances of accidents in different clusters.

The provision of transport infrastructure has grown extensively across the globe through a range of networks of modes which have undergone technological improvements cutting across the motive power, the tracks as well as the means which serve as compartment for passengers and goods. It is also a key player in the transfer and distribution of goods from the input points through the manufacturing line to the customers (Afolabi & O., 2014).

According to traffic accident report provided by World Health Organization (WHO), 1.25 million people die each year on the world's roads and between 20 and 50 million injuries due to road traffic accident (WHO, 2018).

The geographical approach to the study of traffic accidents relates the concept of place, time and environment to accident occurrence. It is believed that land uses, road element, width of the road, bending of road, hilly area, topography and regional distribution in occurrence of road traffic crash are factors to be considered. There are also rural-urban differences. In urban areas, there are more accidents, lower degree of injury while in rural areas, there are lower accident levels but more serious fatalities (Williams, 2015).

The previous researches (Getnet & M., 2009), (Tibebe Beshah & Hill, 2010), (Williams, 2015) and (Addi et al., 2017) have focused merely on higher ways and association rule data mining techniques to identify the interesting rule to understand the circumstances of accident in different cluster, which shows there is a gap for further research to apply other data mining techniques that predict accident casualty and identify the determinant factors contributing to the accident in which J48 and SMO ware applied. Changes on traffic rules and regulations are made in the capital city, which has its own contribution in road safety after these researches have been done. Also Data Mining Algorithms were used to analyze the best algorithm in terms of accuracy and speed in predicting Accident Casualty and a One Factor Prediction Algorithm with a Model ware designed.

3.0 METHODOLOGY

In this paper CRISP-DM standard data mining methodology ware adopted. The important iterative activities that are undertaken in this research are: business understanding, data understanding, data preprocessing and selection of modeling technique, model building and model evaluation. Due to its documentation, inclusion of features to handle almost all activities performed in any data mining methods, and availability of the software; WEKA data mining tool is selected and used for this research.

3.1 Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to familiarized with the data to identify data quality problems, discover first insights into the data, or detect interesting subsets to form hypothesis for hidden information (CRISP-DM, 2000).

Three Domain experts ware consulted to have insight into the problem domain. The domain experts constitute three individuals from FRSC, Police and Drivers that are in charge of road accident. On the basis of the insight gained from discussion with domain experts and review of relevant documents, a clear understanding of the data is achieved.

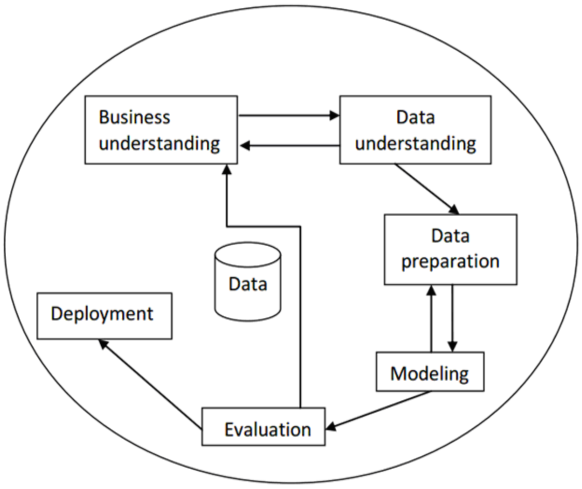


Figure 3.1: Phases of the CRISP-DM reference model (adapted from the CRISP-DM consortium, 2000)

3.2 Data Preparation for Analysis

(a) Data/Attribute Selection

Records are evaluated and classified based on the values of their attributes. Of course, some of the attributes of a record may be irrelevant to the process of classification and thus should be excluded.

Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean (Whitten, 2005). Out of the 32 attributes of the original data set, 14 attributes (including the class attribute) which are believed by the domain experts to have significant contribution in assessing road, drivers and casualty class which is the focus of this research, as shown in Table 3.1.

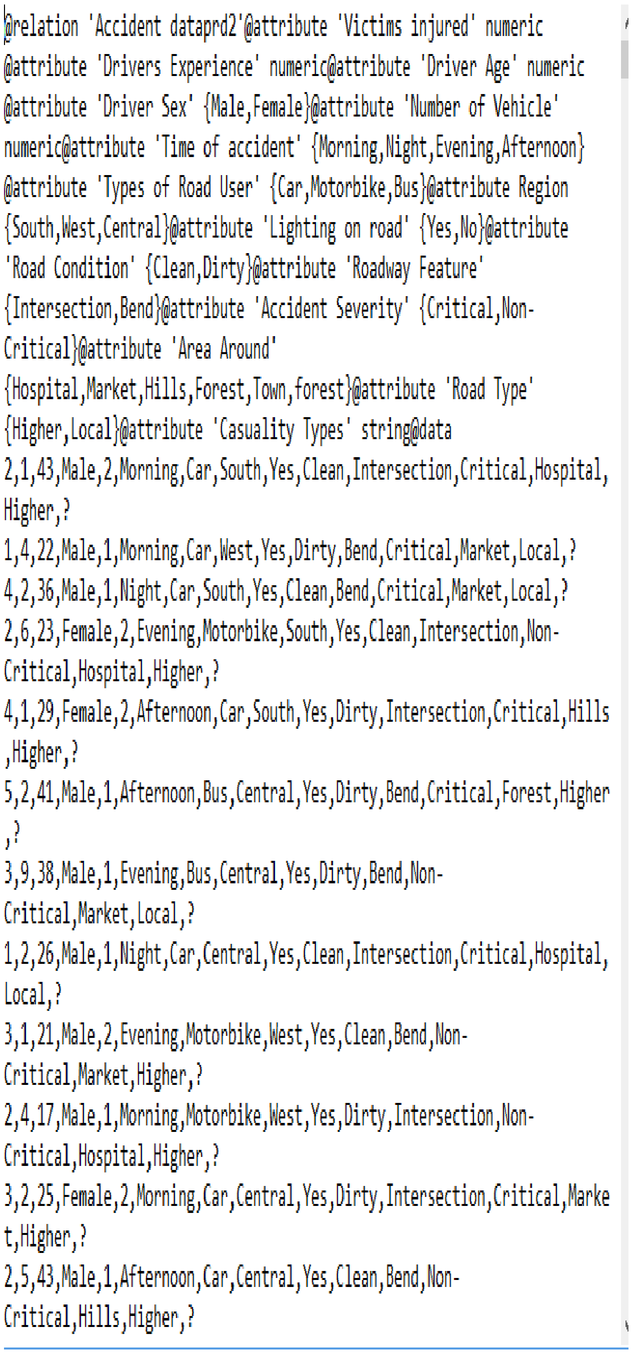
Table 3.1 Description of the whole attributes

S/No.	Attributes	Instances
1.	Victim injured	Numeric
2.	Driver's Experience	Numeric
3.	Driver's Age	Numeric
4.	Driver's Sex	Male, Female
5.	Number of Vehicle collision	Numeric
6.	Accident Time	Morning, Afternoon, Evening, Night
7.	Region	South, Central, West
8.	Lightening on Road	Yes, No
9.	Road Condition	Clean, Dirty
10.	Roadway Features	Intersection, Bend
11.	Accident Severity	Critical, Non-Critical
12.	Area Around	Market, Hospital, Forest, Town
13.	Road Type	Higher, Local
14.	Casualty Type	Fatal, Minor

(b) Data Transformation

Data transformation can involve, smoothing or feature (attribute) construction, which works to remove noise from the data. Smoothing techniques include binning, regression, and clustering. Attribute construction on the other hand is a process where new attributes are constructed and added from the given set of attributes to help the mining process. Smoothing can also serve as data reduction, for example in the case of smoothing through binning; the number of the distinct values for a certain attribute is reduced (Han, 2006).

Data transformation aims to manipulate the data so that its content and its format are most suitable for the data mining process. Accordingly, based on the attribute types of road user to categorize the input values as Car, Motorbike or Bus. Accident_ Time is also derived from Accident Time category attribute to classify the input values as less than Morning, Afternoon, Evening, and Night.



```
@relation 'Accident dataprd2'@attribute 'Victims injured' numeric
@attribute 'Drivers Experience' numeric@attribute 'Driver Age' numeric
@attribute 'Driver Sex' {Male,Female}@attribute 'Number of Vehicle'
numeric@attribute 'Time of accident' {Morning,Night,Evening,Afternoon}
@attribute 'Types of Road User' {Car,Motorbike,Bus}@attribute Region
{South,West,Central}@attribute 'Lighting on road' {Yes,No}@attribute
'Road Condition' {Clean,Dirty}@attribute 'Roadway Feature'
{Intersection,Bend}@attribute 'Accident Severity' {Critical,Non-
Critical}@attribute 'Area Around'
{Hospital,Market,Hills,Forest,Town,forest}@attribute 'Road Type'
{Higher,Local}@attribute 'Casuality Types' string@data
2,1,43,Male,2,Morning,Car,South,Yes,Clean,Intersection,Critical,Hospital,
Higher,?
1,4,22,Male,1,Morning,Car,West,Yes,Dirty,Bend,Critical,Market,Local,?
4,2,36,Male,1,Night,Car,South,Yes,Clean,Bend,Critical,Market,Local,?
2,6,23,Female,2,Evening,Motorbike,South,Yes,Clean,Intersection,Non-
Critical,Hospital,Higher,?
4,1,29,Female,2,Afternoon,Car,South,Yes,Dirty,Intersection,Critical,Hills
,Higher,?
5,2,41,Male,1,Afternoon,Bus,Central,Yes,Dirty,Bend,Critical,Forest,Higher
,?
3,9,38,Male,1,Evening,Bus,Central,Yes,Dirty,Bend,Non-
Critical,Market,Local,?
1,2,26,Male,1,Night,Car,Central,Yes,Clean,Intersection,Critical,Hospital,
Local,?
3,1,21,Male,2,Evening,Motorbike,West,Yes,Clean,Bend,Non-
Critical,Market,Higher,?
2,4,17,Male,1,Morning,Motorbike,West,Yes,Dirty,Intersection,Non-
Critical,Hospital,Higher,?
3,2,25,Female,2,Morning,Car,Central,Yes,Dirty,Intersection,Critical,Marke
t,Higher,?
2,5,43,Male,1,Afternoon,Car,Central,Yes,Clean,Bend,Non-
Critical,Hills,Higher,?
```

(c) Data Set Format

WEKA file starts with the dataset's name followed by list of attributes. In fact, the dataset's name should be preceded by the symbol "@" and the word "relation" (for example; @relation 'Accident data2' is the name of the dataset) and each attribute name also starts with the same symbol and the word "attribute" and following the name of the attribute including its possible values. If the variable or attribute is nominal, a list of possible values contained in a brace is required. By default, the last attribute in the list of the attribute of the dataset designates the target class.

Selection of Modeling Technique

According to the CRISP data mining standard methodology employed in this paper, selecting the actual modeling technique to be used is the first step in modeling (Whitten, 2005)

In order to accomplish this work, two data mining techniques i.e. Decision trees (using J48 algorithm) and Function (using SMO) ware used.

Decision trees are easy to build and understand as shown in figure 3.1. They can handle both continuous and categorical variables and can perform classification as well as regression. It automatically handles interactions between variables and identifies important variables.

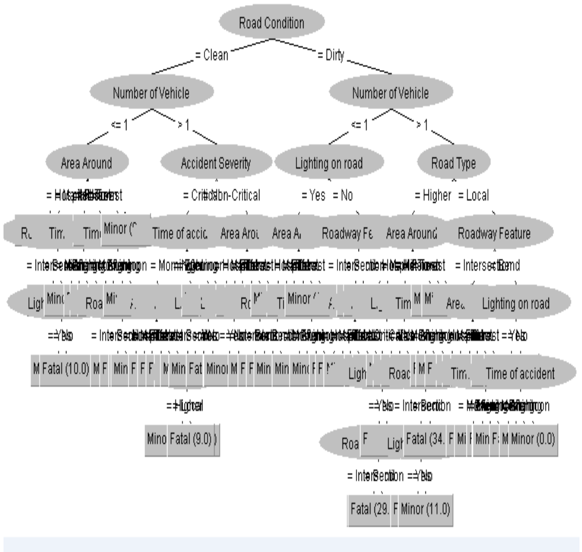


Fig. 3.1: Decision Tree

4.0 RESULTS AND DISCUSSION

4.1 One Factor Prediction Algorithm (OFPA) Experiment

The experiment is conducted using 9 attributes for one factor (Road) that are selected during the data preparation phase. These are Types of Road User, Time of accident, Lightening on Road, Roadway feature, Accident Severity, Area Around, Road Condition, Road Type and Casualty Type. The Casualty Type attribute is the dependent attribute and the rest are independent variables or predictors.

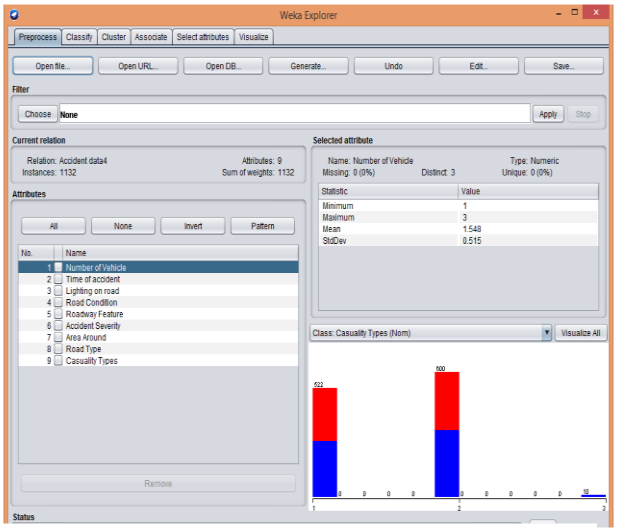
Function: *One Factor Prediction Algorithm (OFPA)*

```
Input: Dataset D;//To be classified
Output: Predicted Values(F,M)
Let A=one factor instances
A= {AA=Hospital&&NV=<=1&&RF=Bend&& LR}
Maxattribute=9
MinInstances=1
for i=1 to Maxattribute-1 do
   for j=1 to Maxattribute
if A==Nonnegative then
PClass=Mp;
   if PClass==Pvalue then
A=true
        else
A =False
End if
End if
End for
End for
```

4.1 Decision Tree Building using J48 Algorithm

WEKA has implementation procedures of numerous classification and prediction algorithms to develop decision tree. J48 algorithm of decision tree technique is one of these algorithms which support both numeric and nominal predicators and nominal class attribute values. J48 algorithm is an implementation of the C4.5 decision tree learner. The algorithm for induction of decision trees uses the greedy search technique to induce decision trees for classification. There are many parameters which can be adjusted in order to obtain better models with respect to the accuracy (or other parameters which can be used as measure for the quality of the model). These parameters allow greater control of the user in the process of learning the models (Whitten, 2005)

J48 has a facility of generating outputs both in tree form and rule sets. The set of rules are generally easier to understand since each rule describes a specific context associated with a class. It also shows the hierarchy of the determinant factors or attributes. In previous section, there is a brief explanation of how J48 algorithm works and description of parameters. The cleaned and preprocessed dataset of csv format is fed to WEKA software. All the selected attributes and other dataset are shown in figure 4.1.



None

Current relation

Relation: Accident data4 Attributes: 9
Instances: 1132 Sum of weights: 1132

Attributes

No.	Name
1	Number of Vehicle
2	Time of accident
3	Lighting on road
4	Road Condition
5	Roadway Feature
6	Accident Severity
7	Area Around
8	Road Type
9	Casuality Types

Selected attribute

Name: Number of Vehicle Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	3
Mean	1.548
StdDev	0.515

Class: Casuality Types (Nom)

Figure 4.1: A screenshot that shows attributes of One Factor

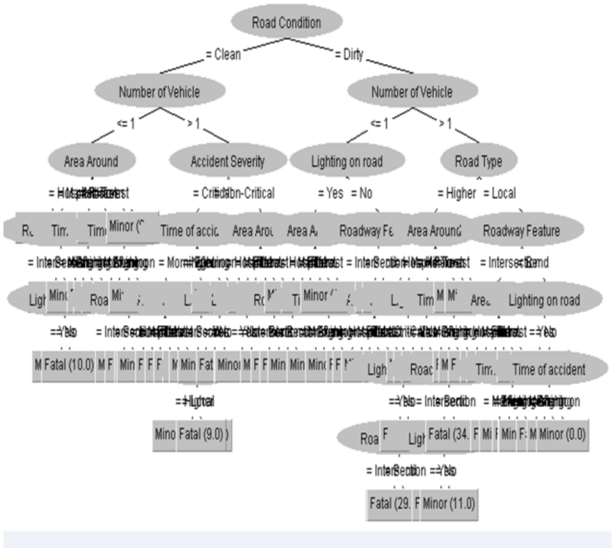


Figure 4.2: J48 Tree for OFPA

In this first experiment all the 9 attributes related to Road Accident, which are believed by subject experts are used to build the model. Since the explorer generally chooses sensible defaults. (Whitten, 2005) the J48 decision tree algorithm with all its default parameters is run on the dataset. The default values for some of the parameters are: 0.25 for the confidence interval, pruning is allowed, the minimum number of objects for a leaf is 2. The training and testing is done using tenfold cross validation.

The k-fold (k=10) cross validation test options is used because the data set has unbalanced number of dependent class values; by doing so the partition and experiment could be more reliable. In this test option the accuracy estimate is the overall number of correct classifications from the k iteration divided by the total number of samples, which is k. After deciding the values of the parameters the algorithm is run to start building the model.

Table 4.1: Statistical Analysis of Classifiers with Cross Validation

Name of Classification Algorithm	Class	TP Rate	FP Rate	Preci-sion	Recall	F-Meas-ure	ROC Area	Accuracy (%)
J48	Fatal	0.952	0.038	0.966	0.952	0.959	0.982	95.67
	Minor	0.962	0.048	0.946	0.962	0.954	0.982	
SMO	Fatal	0.760	0.439	0.660	0.760	0.707	0.660	66.60
	Minor	0.561	0.240	0.675	0.561	0.613	0.660	

In this experiment, the J48 algorithm used Accident Severity and Road Type for one Factor in its pruned

tree in few lines as shown in figure 4.2. This indicates that these attributes are considered as insignificant to discriminate records.

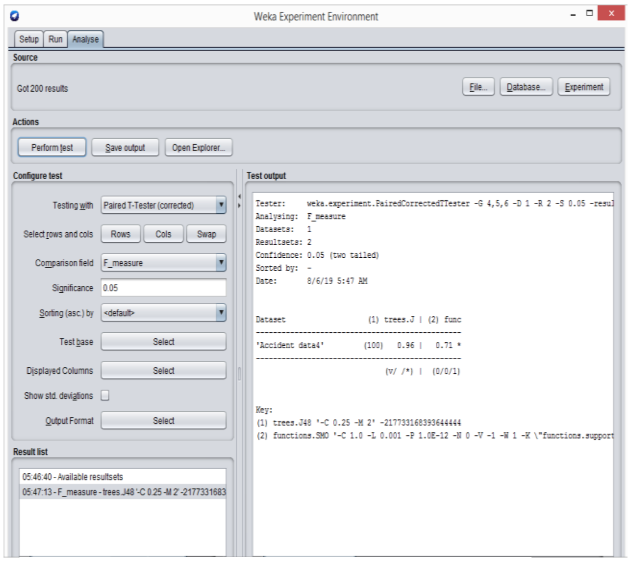
4.2 Models Evaluation

Evaluation is one key point in any data mining process. It serves two purposes: the prediction of how well the final model will work in the future and an integral part of many learning methods, which help to find the model that best represents the training data.

In the series of experiments, evaluation of models is done based on performance/accuracy of models and confusion matrix, discussion with the domain expert and based on the soundness of the rules generated. It is easy to learn that all the two classifiers are performing well and almost similarly with respect to the number of correctly classified instances.

But accuracy by itself doesn't tell everything about the efficiency of predictions, WEKA's experimenter has been utilized to implement j48 and SMO algorithms to automatically analyze the models. The experiment type is a ten-fold cross-validation and model parameters have been set.

The two algorithms are executed at the same time, on the same data set, but not equal number of attributes. Figure 4.3 below, shows the set-up of WEKA's experiment environment window after the data is fed and the two algorithms J48 and SMO have been selected.



Got 200 results

Testing with: Paired T-Tester (corrected)
Select rows and cols
Comparison field: F_measure
Significance: 0.05
Sorting (asc.) by: <default>
Test base
Displayed Columns
Show std. deviations □
Output Format

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -resul
Analysing:  F_measure
Datasets:   1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       8/6/19 5:47 AM


Dataset                   (1) trees.J | (2) func
-------------------------------------------------
'Accident data4'        (100)   0.96 |   0.71 *
-------------------------------------------------
                                (v/ /*) |  (0/0/1)


Key:
(1) trees.J48 '-C 0.25 -M 2' -217733168393644444
(2) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.support
```

```
05:46:40 - Available resultsets
05:47:13 - F_measure - trees.J48 '-C 0.25 -M 2' -2177331683
```

Figure 4.4 WEKA model comparison result

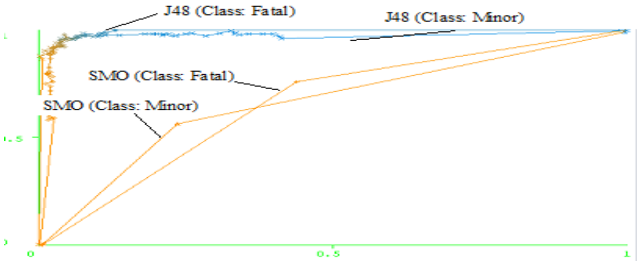


Figure. 4.5: An OFPA Performance Model Chart

After finishing the necessary setup, the algorithms are run to do the experiment. To analyze the experiment, the percent-correct comparison measurement factor is selected from comparison field box as shown in figure 4.4, and then the model has been analyzed by the experimenter automatically. The comparison result showed that the J48 algorithm outperforms than SMO algorithms with an accuracy of 96% and 71% respectively. The Performance Model Chart among the algorithms is shown in the figure 4.5.

4.3 Results

The Result of One Factor Prediction Algorithm (OFPA) as indicate that, out of the 1,132 records of the dataset 1,083 records are correctly classified and the model has an accuracy of 95.67%. The confusion matrix also shows that 570 out of 590 Fatal and 513 out of 542 Minor records are classified correctly. 754 records are correctly classified out of 1,132 with 66.61% accuracy. The confusion matrix also shows that 454 out of 679 Fatal and 308 out of 453 Minor records are classified correctly.

4.4 Discussion

From the experiment J48 algorithm has higher prediction accuracy with 95.67% and using J48 algorithm of decision tree technique, The Road Condition, Number of Vehicle, Area Around, lightening on the Road, Accident Severity and Road Type are the most determinant factors.

5.0 CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In this paper an attempt has been made to apply the J48 decision tree and SMO function predictive data mining techniques in One Factor Predicting Algorithm (OFPA) for car accidents and identify the determinant factors in the accident data set. To achieve this goal: the CRISP-DM 1.0 standard data mining methodology has been adopted and the WEKA data mining tool has been used to implement the algorithms.

The data for this research is the Accident data of the years 20015-2019 collected from FRSC, Police and Drivers. After preprocessing out of 3,841 records, 1,132 Accident records are used for building the models. Various experiments ware made iteratively by adjusting the parameters and used different attributes for comprehensive output. The Major determinant factors that cause road accidents are identified; The Road Condition, Number of Vehicle, Area Around, lightening on the Road, Accident Severity and Road Type are the most determinant factors. Data mining rules ware generated using J48 decision trees and SMO function algorithm and the results were compared. The comparison of the models using WEKA's experimenter showed that J48 outperforms SMO algorithms with an accuracy of 96% against 71%.

5.2 Recommendations

This paper is conducted mainly for academic purpose. However, the researcher highly believes that the findings of this paper can be used by stakeholders to further investigate the nature of road accident problem in Katsina State. The following are recommendations based on the result of this study.

- It is clear that for an efficient data-mining task there is a need for an availability of electronic data. Although the records are good, however stakeholders should take measures to store all the records with all the necessary attributes in an electronic format. To reduce this misclassification error, the offices has to Re-engineer the data collection system, train the data entry personnel, and fully automate the data Recording system in collaboration with other organizations like WHO and NBS. This will enable the offices to have a quality data which in turn increases the accuracy of the models to be built.

Predictive models that do not require apriori output variable, such as association rules, should be further investigated. However, the model shows a good level of accuracy of 96% it should not be the only parameter to consider, another future work is to test the applicability of other data mining techniques on other regions of the country.

REFERENCES

Addi, A.-M., Tarik, A., & Fatima, G. (2016). An Approach Based On Association Rules Mining To Improve Road Safety In Morocco. *IEEE*.

Adekunle, J. A. (2012). Road Traffic Accident Injuries and Productivity in Nigeria. Journal of Asian Scientific Research, 2(7), 334-344.

Chong, M., Abraham, A., & Paprzycki, M. (2004). Traffic accident analysis using decision trees and neural networks. In: Isaias P et al (eds) IADIS International Conference on Applied Computing.IADIS Press Portugal, 2, 39-42.

CRISP-DM, T. (2000). Step-by-step data mining guide available at www.crisp.org

Getnet, & M. (2009). Applying data mining with decision tree and rule induction techniques to identify determinant factors of drivers and vehicles in support of reducing and controlling

Gwamki, U. B., Liman, A., & Rumah, S. G. (2018). The State and the Management of Road Traffic Crashes in Nigeria: The Case of Federal Road Safety Corps (FRSC), Katsina. International Journal of Social Sciences and Management Research, *4*(5), 24-36.

Han, Jiawei, Kamber, & Micheline. (2006). Data Mining: concepts and Techniques. Morgan kufman

Kuhnert, P., Doo, K., & McClure, R. (2000). Combining nonparametric models with logistic regression: an application to motor vehicle injury data. . *Comput Stat Data Anal*, *34*(3), 371-386.

Sachin, K., & Durga, T. (2015). Analysing Road Accident Data Using Association Rule Mining. *IEEE*, *1*, 978.

Tibebe Beshah, & Hill, S. (2010). Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia.

Whitten, I., H., Frank, & E. (2005). Data Mining: practical machine learning tools and techniques with java implementations. *Morgan Kaufmann*

Williams, M. E. (2015). *Analysis of road traffic crash along lokoja – abuja highway, nigeria.* (masters degree of science in geography), ahmadu bello university, nigeria.

Zhang, M., Kecojevic, V., & Komljenovic, D. (2014). Investigation of haul truck-related fatal accidents in surface mining using fault tree analysis. *Saf Sci*, *65*, 106-117.

DESIGN OF VOICE RECORDER IN THE ANDROID PLATFORM

[1]Bunakiye R. J. and [2]Iniakpo T.
[1,2]Department of Mathematics/Computer Science, Niger Delta University Wilberforce Island, Bayelsa State, Nigeria
*Email of Corresponding Author: bunakiye.japheth@ndu.edu.ng

ABSTRACT

The paper is based on the design of an android application that provides easy recordings to users like reporters, musicians, and students on a mobile device. The number of libraries in android in combination with the SDK and Java run time (JRE) provided the support for the development of this application. By this developed application even messages can be sent via a voice command and can also retrieve unread messages in form of voice notes. It is designed to aid recording of sounds for multi-users, and embedded with newer features, smarter interface, and quicker navigation and enhanced functionality to enable usage of the voice recording facility by multiple mobile users. The process implementation logic followed human computer interaction analysis and design using the android studio and Unified Modeling Language (UML). The design continues defining the collaborations between the different objects where the requirement of the primitive elements are used to glue together the classes and the procedures. The goal therefore is to provide an easy way for the users to record voice for later or proper use through a mobile app. This access to user comments just by a voice command is all possible with the built-in application programmer interfaces.

Key Words: Android Studio, Multi-User Environment, Digital Sound Recorder, Audio Controller, System Subclasses

1.0 INTRODUCTION

The focus is on designing a voice recorder capable of recording sounds particularly in the android environment. The essence is that newer features such as a smarter interface, and enhanced navigation functionalities are added to enable usage of the voice recording facility by multiple mobile users [20]. The applicability in this paper is leveraging the capacity of the android platform to bringing to bear recording creativity onto a mobile device. In order to achieve this progress, the characteristic work flow capabilities of the android platform are deployed to develop an android application, which helps in recording sounds in the sense of remote applicability with clear and easy data access [14]. Again exploiting the functionalities of a mobile device by making all features available in the application is about bringing to bear the requirement engineering of the primitive elements necessary for a system of this kind to be developed. This is therefore more of a human computer interaction platform with the aid of recording sounds in a particular environment encompassing newer features, smarter interface, quicker navigation and enhanced functionalities for purely confidential matters [10].

Considerably, the use of jotters and cassette recorders has become globally accepted means of recording both written and voice data but in some cases where secret investigations and undercover works are involved; these means of voice data collection have posed several data detection problems because of the kind of operational platforms associated with them. The android mobile platform as the case may be emits network frequencies, which makes it very hard for a third party to detect if someone is recording a conversation or not [13]. Android has a number of libraries to provide support for development of the android application. Android mobile application is responsible for executing stored commands. For developing this Android application, Android SDK and Java run time (JRE) were used. By this developed application even messages can be sent via a voice command and can also retrieve unread messages in form of voice notes. In many identified situations, this mobile voice recording system can be of immense support to music makers by helping them to keep a voce record of any musical gimmick before proceeding to a studio for proper recording.

In addition this application offers ease of voice interaction between users even if it involves paying bills without necessarily getting connected through via a website. The inclusive implication [11] is to bring the ‘Recorder Functionality’ onto a mobile device that can record sound in a distinct coverage area with clear audibility and can as well make it easy for users to gain access to recorded data. This effort actually exploited the functionalities of a mobile phone by making all features available in on platform thus meeting the requirements to a rising need for simplicity. The presence of all sound recording related activities in one application is customer friendly and by implication takes mobility a step higher [22]. Being installed on a portable mobile device, the user will not have to change cassettes or go to the studios to perform activities that can simply be completed with a touch on the screen of the device in their own location.

2.0 RELATED WORKS

This paper is basically presenting an android application that provides easy recordings to users such as engineers, reporters, musicians, students, and medical practitioners etc. The goal is to bring the functionalities of a recorder onto a mobile device. The android platform with the open source capabilities, supportive multifunction, and rich tools provision for interactive application development was selected to implement this app. Santhrushna Gande [1], posited more on the wide acceptance and usage of android operating system based mobile phones and tablets around the world. The popularity according to [3] of this operating system is due to its multi-tasking, ease of access and diverse device options. For example the Java programming speech recognition system is an android application used for handicapped individuals who are not able to type on a keyboard. This application by virtue of its application logic enables the user to write a computer program in Java by dictating the words and without using a keyboard. This work compliments ours in the utilization of the voice recording creativity that enable user's record voice data for customized usage.

Yousra Aafer et al. [5] in their work on hanging attribute references characterized android by the complicated relations among its components for application development. They further designed and implemented Hare hunter, a new tool for automatic detection of Hares by comparing attributes defined with those used, and analyzing the references to undefined attributes to determine whether they have been protected through appropriate checks. The process was such that the conflict between the decentralized and the interdependency among different Android components, a type of vulnerabilities which often have serious security implications were solved. This work did showcase components selection and attributes description such that the required functionalities are met.

Renu Tarneja1 et al. [2] presented a realization where smart mobile android devices have become so portable that one device is an embodiment of functionalities such as a smart phone, a music player, a notepad, a navigation unit and lots more. Even though today's smart phones do so many new things, voice recognition, which is able to recognize human speech and change to text, or to perform a command, has emerged as the next most fascinating thing possible with android devices. Voice actions are a series of spoken commands that let the user control the mobile device with his or her voice; a technology that uses desired equipment and a service which can be controlled through voice without touching the screen of the android smart phone. Consequently, these researchers having taken into cognizance that speech recognition is a hard problem, requiring a combination of many techniques; introduced through modern methods some new set of features for voice actions for android platform smart mobile devices. Their proposed device is an interactive android smart phone, which is capable of recognizing spoken words, help the user to open any application, and can as well call any contact through voice. Similarly, our voice recorder system operating on the same android platform help the user to achieve virtually all the interactive components using the voice component functionality.

Characterizing smartphone usage patterns from millions of android users Li et al. [18] presents an empirical analysis of app usage behaviors collected from millions of users that bothered on Android apps in terms of app management activities (i.e., installation, updating, un-installation and updating of app network traffic. In their work multiple aspects of such behavior data were explored and interesting patterns of app usage present. The results provide many useful implications to the developers, users, and disseminators of mobile apps. The remarkable difference with this paper is in the aspect of research direction. Central to this research also is the analysis of functionalities of the android platform for application development and the subsequent follow up activities that led to the development of a multifunctional sound recorder suitable for mobile devices. William Enck et al. [17] seeks to better understand smart phone application security by studying several popular free Android applications. Their analysis uncovered misuse of phone identifiers, and deep penetration of advertising and analytics networks. Though they considered the security implications through code discovery techniques, no evidence of malware or exploitable vulnerabilities in the studied applications were discovered. Understanding the characteristics of android wear, a popular wearable OS, Renju and Xiaozhu [6] showcased how interactive wearable devices bring dramatic changes to apps and hardware, leaving operating system (OS) design in the mist. They examined the execution efficiency of Android Wear by running a suite of benchmarks, and profiling CPU usage, idle episodes, thread level parallelism, and micro architectural behaviors. These are system aspects, which we also considered in this current design.

Onukem and Oye [19], discussed the relevance of feature-rich android SDKs in the development of ubiquitous applications in smart phones, tablets and other mobile devices. They compared a number of cross-platform frameworks for mostly platform-independent development on mobile platforms, by evaluating them in several categories and weighing them against native SDKs. In the same manner cross-platform solutions were identified but were found to be limited in high requirements applicable to performance, and usability experience. Therefor to meet with the increasing demand for mobile applications actions many areas especially in voice actions, the android platform was chosen for implementation. One influencing factor is the android lead in the sales with a high market share, while other platforms were significantly less prevalent in the development circumstances.

3.0 METHODOLOGY

Users are increasingly turning to their mobile devices when doing most of their human interactions. Although mobile device usability has improved, some usage scenarios involving typing can still be cumbersome, error-prone, and even risky. The goal therefore is to provide an easy way for the users to record voice for later or proper use through a mobile app. This is all possible with the built-in application programmer interfaces. The app can access user comments just by a voice command. Before any work can begin on the development of an android application, the first step is to configure a computer system to act as the development platform [19].

This involves a number of steps consisting of installing the Java Development Kit (JDK) and the Android Studio Integrated Development Environment (IDE), which also includes the Android Software Development Kit (SDK) [21].This section will cover the conceptual approach taken to actualize the necessary requisite components installation and the application development implementation. Since the Android SDK was developed using the Java programming language, which also means that android applications are also developed using Java. As a result, the Java Development Kit (JDK) is the first component that must be installed. Android Studio 2 development requires the installation of version 8 of the Standard Edition of the Java Platform Development Kit [12]. Java is provided in both development (JDK) and runtime (JRE) packages. For the purposes of Android development, the JDK must be installed. For Windows systems, the JDK can be obtained from Oracle Corporation.

Once downloaded, the installation executable is launched and the on screen instructions critically followed to complete the installation process. Most of the work involved in developing applications for Android will be performed using the Android Studio environment [2]. The content and examples in this paper were created based on Android Studio version 2.2. The downloaded Android Studio installation executable file (named android-studio-bundle<version>.exe) is located in the Windows Explorer window to start the installation process. Once the Android Studio setup wizard appears, the various screens are thoroughly examined to configure the installation to meet the current requirements in terms of the file system location into which Android Studio should be installed [5].

The requirements selection process starts with the installation components making sure that the Android Studio, Android SDK and Android Virtual Device options are all selected [17]. Although there are no strict rules on where Android Studio should be installed on the system, the present assumption is that the installation was performed into C:\Program Files\Android\Android Studio and that the Android SDK packages have been installed into the user's AppData\Local\Android\sdk sub-folder [4]. On versions of Windows with a Start menu, the newly installed Android Studio can be launched from the entry added to that menu during the installation. The executable may be pinned to the task bar for easy access by navigating to the Android Studio\bin directory, right-clicking on the executable and selecting the Pin to Taskbar menu option [20]. To note at this point is that the executable is provided in 32-bit (studio) and 64-bit (studio64) executable versions. So the user need to be sure to use the exact studio executable running on the computing machine.

4.0 DESIGN

This section describes the object-oriented design and implementation of the sound recorder, using UML notation. There are several digital sound recorders commercially available in the market. The model described here has been designed following the specification requirements of the object model of the system as the main class diagram [9]. The internal behaviour of each object show the hardware architecture of the sound recorder and the concurrency model, where each object is assigned to an execution thread [15]. The Use Cases describe the functionality of the system from the user's point of view. The user may be a person or another machine. Each use case is a different way to use the system and the completion of each use case produces a different result. In this system there are six different use cases, represented in the Figure 1

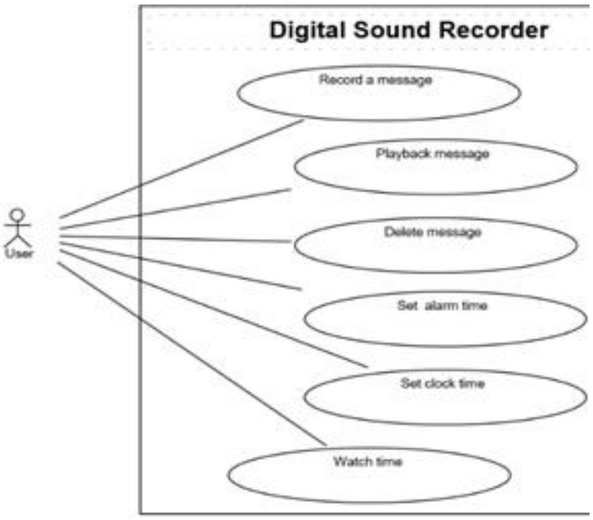


Figure 1: Use-Case Diagram for the Application

The user interacts with the system through the display and the keyboard. These are somehow passive objects, so a user interface object is added to manage the interaction with the user. The User Interface class relies on the Audio Controller to perform the tasks [8]. The Audio Controller is the core of the system and it executes the tasks suggested by the user interface. The audio controller uses the Audio Input and Audio Output classes to produce the sounds. These classes are related with a microphone and speaker hardware as shown in Figure 2.

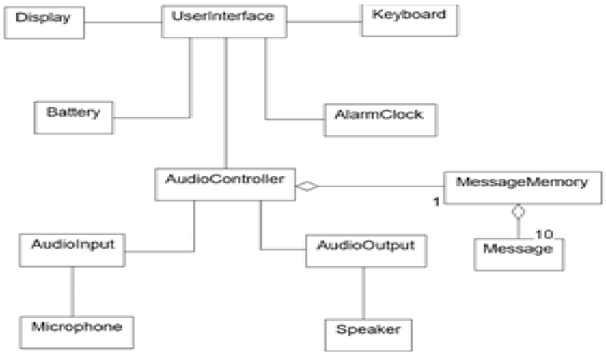


Figure 2: User Interaction Class Diagram

A user interaction class diagram presents a static view of our system. The object diagram in Figure 3 is used to depict the dynamic behaviour of the system and to understand the internal evolution of a single object. Audio Controller object is the reference point for the sound channels [18]. A sound channel can be used to record a message, to play a message or to play an alarm sound.

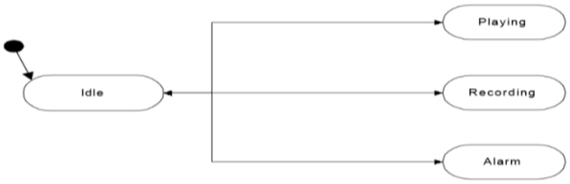


Figure 3: The Object Class Diagram

Android is designed primarily for mobile devices such as smartphones and tablet computers, and for this reason [6], an algorithm is designed to depict the deterministic event handler in the voice action scenarios, which invariably is to enhance the software reliability.

```
Input: Voice Input; Voice Output
voice stream
procedure VOICE RECOGNITION()
while processing
executable model; true
locator   STREAMING SIGNALS()
if signal is null then
source  GETJRECODENEXT()
executable   CHECKSOURCECODE(voice signal)
end if
execute signal if executable is true
end while
end procedure
procedure CHECKSYNTAX
signal   PROCESSINGMODEL()
return null
end if
if signal times out
end if
return semantics event.signal
end if
end procedure
return true
end if
end if
end procedure
```

Figure 4 typically represents the attributable voice data input recognition and output processing procedure to enable the user actualize the usable scenario.

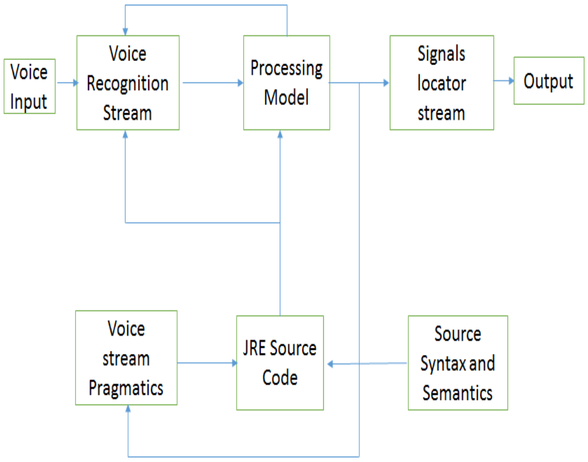


Figure 1. Typical Voice Processing Procedure

The procedures involves the voice recording attributes and their relationships [17]. This attributes are characteristics of a vocabulary that is used both by the android platform logic and the users. The platform logic as shown in Figure 5 represent the things, ideas and concepts that are contained in the application.

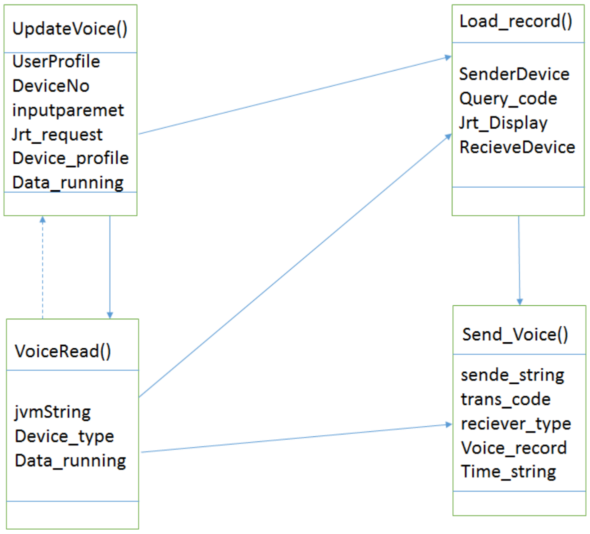


Figure 5: Platform Attributes

The attributes display the containment, inheritance, and the association relationships that lead to implementing the usable scenarios. Significantly, the arrows simply divide into three compartments. The topmost compartment contains the load and update classes [5]. The middle compartment contains the member variables list, and the bottom compartment contains the mechanics of the Java run time string operations with the associated member functions. The type of variable arguments or function arguments required to facilitate the member function procedure are shown in Figure 6.

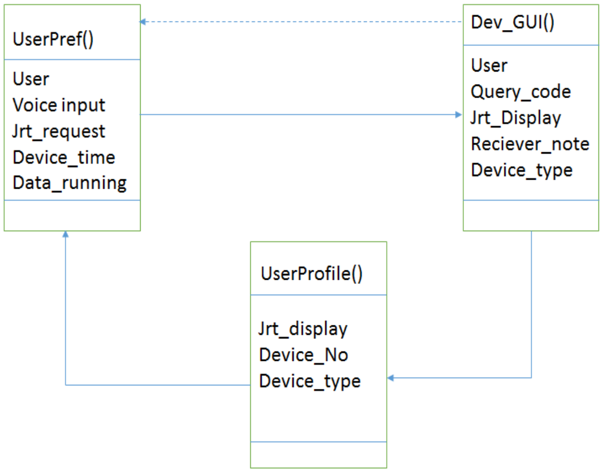
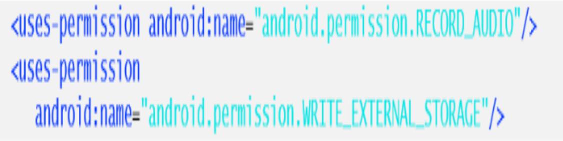


Figure 6: Variable and Function Argument Procedure

The implication for these arguments is to show only those attributes and operations that are useful for the particular scenario during the voice action input and output operations within the platform [16].

RESULTS AND DISCUSSION

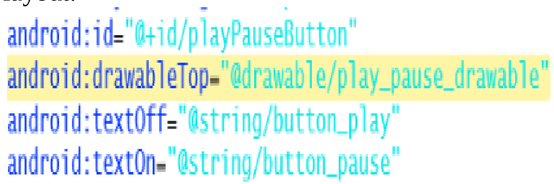
The most fundamental component of the voice recording process is the task that the user records sounds using the phone's microphone and save the audio files for playback later. The activity of the recorder continues after the user finishes a recording. The possible permissible syntax design is:



```
<uses-permission android:name="android.permission.RECORD_AUDIO"/>
<uses-permission
   android:name="android.permission.WRITE_EXTERNAL_STORAGE"/>
```

With this syntax a View Saved Recordings button can allows the user to view a list of saved recordings. When the user touches the Record Toggle-Button, it becomes a Stop Toggle-Button. When the user touches the Stop Toggle-Button, the Save and Delete buttons are enabled so the user can choose whether to save or delete the temporary recording file [13]. If the user touches Save, the Name (i.e. the Recording dialog is displayed. If the user touches Delete, a confirmation dialog is displayed before the recording is deleted.

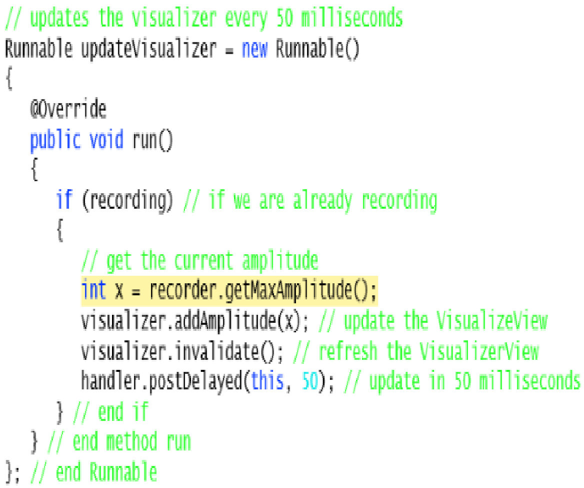
The component activities are specified in the AndroidManifest.xml attribute file, which contains <activity> elements for each Activity such as Voice-Recorder and Saved-Recordings. Both use the portrait screen orientation. In addition, the <manifest> element contains the <uses-permission> elements, which indicate that the app requires the ability to record audio and write data to external storage, respectively [12]. The specification for the layout of the Saved-Recordings-List-Activity provide a List-View with its android: id set to "@android:id/list" when customizing a ListActivity's layout:



```
android:id="@+id/playPauseButton"
android:drawableTop="@drawable/play_pause_drawable"
android:textOff="@string/button_play"
android:textOn="@string/button_pause"
```

This layout introduces the *android:drawableTop* attribute, the custom *Drawableplay_pause_drawable* new feature specified for its *Toggle-Button.* This *Drawable* allows Android to toggle the icon between a play and a pause icon when the user changes the Toggle-Button's state. Drawable for the Play/Pause Toggle-Buttons provide *android:textOff* and *android:textOn* attributes that allow the user to specify the text that is displayed for Toggle-Button's two states [15]. Different icons for the two states can also be specified by defining a custom Drawable that specifies the two icons.

The system consists of an Activity subclass (Voice-Recorder,), a View subclass, (Visualizer-View) and a List Activity subclass (Saved-Recordings). Voice-Recorder is the main Activity class and is responsible for creating a recording and visualizing it. Voice-Recorder also enables the user to save or delete the new recording and to view a separate Activity for playing back previously saved recordings [20]. The only new class used by class Voice-Recorder is Media-Recorder, which declares the VisualizerView that is used to display the visual representation of the audio input while recording. *RunnableupdateVisualizer* updates the *VisualizerView* to reflect the current audio input. If the app is recording, class *MediaRecorder'sgetMaxAmplitude* method is called to get the maximum recording amplitude since *getMaxAmplitude* was last called. We add that amplitude value to the VisualizerView, then call it's invalidate method to indicate that the View needs to be redrawn. It can then schedule *updateVisualizer* to run again after a 50-millisecond delay.



```
// updates the visualizer every 50 milliseconds
Runnable updateVisualizer = new Runnable()
{
   @Override
   public void run()
   {
      if (recording) // if we are already recording
      {
         // get the current amplitude
         int x = recorder.getMaxAmplitude();
         visualizer.addAmplitude(x); // update the VisualizeView
         visualizer.invalidate(); // refresh the VisualizerView
         handler.postDelayed(this, 50); // update in 50 milliseconds
      } // end if
   } // end method run
}; // end Runnable
```

This VisualizerView class is a custom View that displays a visual representation of a recording. As the user makes a recording, the View displays green lines with their height proportional to the amplitude of the audio input. The Saved-Recordings subclass of List-Activity displays the user's saved recordings in a List-View. The user can touch a recording's name to play it, touch the corresponding e-mail icon to send the recording as an e-mail attachment, or touch the delete icon to delete the recording from the device's storage [21]. The user can also toggle between pausing and playing the selected recording, and can adjust the playback position by using import statements.

5.0 CONCLUSION

This voice recorder is capable of recording sounds using the device's microphone, save the recordings for playback later, delete the recordings and as well send the recordings as email attachments. To enable recording and saving files; the human computer interaction logic with the specified permissions processes the audio recording and then writes to the device's external storage. It is basically an android application that provides easy recordings to users like reporters, musicians, and students on a mobile device. The documentation followed the object-oriented analysis and design of human computer interaction system using the android studio and Unified Modeling Language (UML).

The result is a neat and responsive user interface for quality sound recording with the ability for mobile usage. This access to user comments just by a voice command is all possible with the built-in application programmer interfaces, and the necessary collaborations between the different objects where the requirement of the primitive elements are used to glue together the classes and the procedures. The goal therefore is to provide an easy way for the users to record voice for later or proper use through a mobile app.

REFRENCES

[1] Santhrushna Gande (2015), Developing Java Programs On Android Mobile Phones Using

Speech Recognition *california state university - san bernardino*, 004580341@coyote.csusb.edu

[2] Renu Tarneja1, Huma Khan, R. A. Agrawal, Dinesh. D. Patil (2015) Voice Commands Control Recognition Android Apps; International Journal of Engineering Research and General Science Volume 3, Issue 2, March-April, 2015

[3] Enck, W., Ongtang, M., and Mcdaniel, P. (2009). Understanding Android Security. IEEE Security & Privacy Magazine 7,1(January/February 2009), 50–57.

[4] Ware, M. S., and Fox, C. J.(2008). Securing Java Code: Heuristics and an Evaluation of Static Analysis Tools. In Proceedings of the Workshop on Static Analysis (SAW)(2008).

[5] Yousra Aafer , Nan Zhang, Zhongwen Zhang , Xiao Zhang , Kai Chen, XiaoFeng Wang, Xiaoyong Zhou , Wenliang Du, and Michael G. (2015). Hare Hunting in the Wild Android: A Study on the Threat of Hanging Attribute References 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS 2015 - Denver, United State

[6] Renju L., and Xiaozhu L. (2016). Understanding the Characteristics of Android Wear OS MobiSys'16, June 25 - 30, 2016, Singapore, Publication rights licensed to ACM. ISBN 978-1-4503-4269-8/16/06. . .

[7] Pesterev, A., Zeldovich, N., and morris, R. T. (2010). Locating cache performance bottlenecks using data profiling.InProceedings of the 5th European Conference on ComputerSystems(New York, NY, USA, 2010), EuroSys '10, ACM,pp. 335–348.

[8] Ravindranath, L., Padhye, J., Agarwal, S., Mahajan, R., Obermiller, I., and Shayandeh, S. (2012). Appinsight: Mobile app performance monitoring in the wild.InProceedings of the 10th USENIX Conference onOperating Systems Design and Implementation(Berkeley,CA, USA, 2012), OSDI'12, USENIX Association,pp. 107–120.

[9] Shreya Dugar, Harsh Gupta; A Multi-Purpose Voice Controlled System On Android platform *Iosr Journal Of Computer Engineering (Iosr-Jce) E-Issn: 2278-0661,p-issn: 2278-8727, volume 16, issue 5, ver. Vi (sep-oct. 2014), pp 62-68 www.iosrjournals.org*

[10] Okediran O. O., Arulogun O. T., Ganiyu R. A, Oyeleye C. A.Mobile Operating Systems And Application Development Platforms: A Survey International. Journal. Advanced Networking and Applications Volume: 6 Issue: 1: 2195-2201 (2014) issn : 0975-0290

[11] Yang, T., Liu, T., Berger, E. D., Kaplan, S. F., and Moss J. E. B. (2008). Redline: First class support for interactivity in commodity operating systems. In Proc. USENIX Conf. Operating Systems Design and Implementation (OSDI) (Berkeley, CA, USA, 2008), OSDI'08, USENIX Association, pp. 73–86.

[12] M. Bohmer, B. Hecht, J. Schoning, A. Kruger, and G. Bauer.(2011). Falling asleep with angry birds, Facebook and Kindle: a large scale study on mobile application usage. In Proc. of MobileHCI, pages 47–56, 2011.

[13] M. Bohmer and A. Kruger. (2013). A study on icon arrangement by smart phone users. In Proc. of CHI, pages 2137–2146,2013.

[14] N. Chen, J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang. (2014). AR-miner: mining informative reviews for developers from mobile app marketplace. In Proc. of ICSE, pages 767–778,2014.

[15] T. M. T. Do and D. Gatica-Perez.(2014). Where and what: Using smart phones to predict next locations and applications in daily life .Pervasive and Mobile Computing, 12:79–91,2014.

[16] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin. (2010). A first look at traffic on smart phones. In Proc. of IMC, pages 281–287, 2010.

[17] William Enck, Damien Octeau, Patrick McDaniel, and Swarat Chaudhuri (2014). A Study of Android Application Security Systems and Internet Infrastructure Security Laboratory Department of Computer Science and Engineering The Pennsylvania State University

[18] Huoran Li, Xuan Lu, Xuanzhe Liu. (2015). Characterizing Smartphone Usage Patterns from Millions of Android Users IMC'15, October 28–30, 2015, Tokyo, Japan. 2015 ACM. ISBN 978-1-4503-3848-6/15/10 ...$15.00. DOI: http://dx.doi.org/10.1145/2815675.2815686.

[19] Onukem Chijioke, Oye, N. D. Cross Platform Mobile Phone Application Using J-Query Mobile (A Case Study Of A Simple Mobile Chat Application) International Journal Of Computer Science And Mobile Computing A Monthly Journal of Computer Science And Information Technology ISSN 2320–088x impact factor: 6.017 *ijcsmc, vol. 7, issue. 2, february 2018, pg.56 – 72*

[20] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America, 87(4):1738{1752, 1990.

[21] Müller, Roland M., Bjorn Kijl, and Josef KJ Martens. "A comparison of inter-organizational business models of mobile App Stores: there is more than open vs. closed." *Journal of theoretical and applied electronic commerce research* 6.2 (2011): 63-76.

[22] Ms. Anuja Jadhav Prof. Arvind Patil." Android Speech to Text Converter for SMS Application" IOSR Journal of Engineering Mar. 2012, Vol. 2(3) pp: 420-423.

[23] Jarng, S.S.: "Analysis of HMM Voice Recognition Algorithm." Journal of Advanced Engineering and Technology , publication year-2010 volume no. 3(3), page no. 241–249.

AN OWL BASED ONTOLOGY MODEL FOR CLIMATIC CONDITIONS KNOWLEDGE ON MAIZE CROP FARMING: SCENARIO FOR ENHANCING OWL' OBJECT PROPERTY FOR INTELLIGENT SYSTEMS

***[1]Aminu E. F., [2]Oyefolahan I. O., [3]Abdullahi M. B., [4]Salaudeen M. T., [5]Abdulsalam T. A.**
[1,3,5]Department of Computer Science, Federal University of Technology, Minna, Nigeria
[2]Department of Information and Media Technology, Federal University of Technology, Minna, Nigeria
[4]Department of Crop Production, Federal University of Technology, Minna, Nigeria

*Email of Corresponding Author: *enesifa@futminna.edu.ng*

ABSTRACT

The exponential growths of data in heterogeneous forms cut across all human endeavors and disciplines, agriculture for instance. Accessing knowledge in respect to climatic conditions that affects maize crop during planting stage is very significant in order to boost and maintain the crop's maximum yields. However, retrieving or accessing the relevant knowledge to a user's query intension becomes an issue. Therefore, a promising solution towards mitigating this challenge of retrieving relevant information as a result of natural language ambiguity is by modeling data ontologically. Ontology is a data modeling technique for knowledge representation in a machine understandable format. To this end, this paper aims to model an OWL-based ontology for climatic conditions knowledge affecting maize crop during planting stage and enhance the object properties of the concepts in terms of synonyms by using hybridization of Fox-Gruninger, Methontology and FAO-Based ontology development methodologies and written using OWL2 Web Ontology Language RDF/XML syntax. The correctness of the ontology's content and correctness of the ontology development have been constantly validated by the domain experts and via experiments. Thus, the proposed JENA based system provided a relevant knowledge based on user's queries of the subject matter in a more accurate and timely information.

Keywords: Ontology, Object Properties, Climatic Condition, Information Retrieval, Maize

1.0 INTRODUCTION

The existing Web is becoming more interesting as a result of its capability to provide access and platforms to different data repositories in different formats and sizes. This is a good development on one hand because of availability of information but also poses inherent challenge on the other hand because, the accuracy of (relevant) information retrieval becomes a serious research issues. According to literature, this shortcoming is a fall out of lack of cutting edge technology to develop structured data repositories (Iqbal, et al., 2013) and also ambiguities of the natural languages in terms of synonyms, polysemy, meronyms, hyponyms and the likes. However, in order to solve the problem; semantic web is promising.

Semantic web, according to the inventor is not a parallel web to the existing one but meant to gradually evolve from it (Antoniou and Harmelen, 2004). Semantic web or Web 3.0 is an extension of the World Wide Web (W3C) in which documents are filled by annotation in machine understandable markup languages (Choudhury, 2014). As a matter of facts from literature (Sanchez, *et al*., 2011), semantic web is one hand ontology and on the other hand information retrieval. In other words, ontology is most often described as the stronghold of semantic web. In recent research time, the development of ontology and semantic search has attracted so many research attentions. The cutting edge technology – ontology has proven well to address the problem of unstructured data repositories and issues of search as a result of word mismatch.

Different definitions have been given to ontology as a derived field in Artificial Intelligent (AI) but the most popular one is that Gruber (1993) which states ontology as a formal and explicit specification of a shared conceptualization. In this definition, "formal" refers to machine-readable, "shared" refers to agree upon by a group and "conceptualization" is what defines an abstract model describing a particular field of knowledge. Similarly, ontology represent a description of knowledge level or conceptual specification that describe a domain knowledge in a manner that is independent of epistemic state and state of the world (Zacek, 2017). The domain knowledge in this context refers to any real life scenarios. For example; agriculture, medicine, arts and religions. Agriculture largely depends on climate. Hence, climatic factors such as precipitation, solar radiation, wind, temperature, relative humidity all these factors determine distribution of crops and their productivity (Sokoto, *et al*., 2016).

In most regions with rain fed or irrigation crops, there is highly dependency on precipitation during peak month of precipitation (Neenu, *et al*., 2013). Too much precipitation can cause disease infestation in crops, while too little can be detrimental to crop yields, especially if dry periods occur during critical development stages. Moisture stress during the flowering, pollination, and grain-filling stages is especially harmful to maize (Kidane, 2018). Maize or corn (*zea* mays L.) is an important annual cereal crop of the world belonging to family Poaceae. The term "zea" is an ancient Greek world with meaning "sustaining life" and "mays" is a word from Taino language meaning "life giver". Maize production is highly influence by climatic condition which includes water, light, temperature, relative humidity, air and wind. In this paper, we identified three major seasons during which maize crop are affected by climatic conditions. namely: pre-planting season where maize's seeds are affected, planting season where maize plants are affected and post planting where maize (grains) harvest are affected. However, this work only considers the planting season.

Ontology development includes activities to be carried out, techniques used and steps involved for developing ontology, clarification of point related to the formalisms, tools and language to be use in the development process and the likes (Slimani, 2015). In order to develop ontology just like any other systems a set of guideline called methodologies have to be clearly defined either by adoption or by adaption. These ontology methodologies include Methontology, Uschold and King, Noy and Mc Guinnies, Gruninger and Fox, Tove and the likes (da Silva, 2012). Ontology languages are used to encode ontology as knowledge representation language. There are some bases for selecting any knowledge representation language either expressivity or formality. Ontology languages includes DAML+OIL, Ontology interchange language (OIL), Web Ontology Language (OWL), Resource description framework (RDF), RDF Schema (RDFS), Extensible Markup Language(XML). Ontology tools use in creation and manipulation of ontology are Protégé, OntoEdit, WebODE, WebOnto Swoop and the likes (Alatrish, 2013). OWL provides a means for describing properties and concept in a human and machine-understandable format, Antonius and Harmelen (2009).

Therefore, the proposed paper aims to model climatic conditions knowledge for maize crop farming ontologically. This is by taken into account the relevance of OWL's relations (object property) and its inherent effects of synonyms. For example, the term *affects* is a relation that connects concepts *Heavy Rainfall* and *Maize Crop*. Similarly, the term *influences* is also encoded as a synonyms of relation affects.

2.0 LITERATURE REVIEW

Web Ontology Language (OWL) amongst others is one of the foremost ontology knowledge representation languages. Currently, aside from the existing OWL1, it has improved upon and thereby gave birth to OWL2 in 2012. OWL provides sophisticated modelling constraints such as explicit cardinalities, universally and existentially quantified property constraints, and class definitions based on the union, intersection, or complement of other classes. OWL consists of namespace, classes, properties, and individuals (Cardoso and Pinto, 2015). OWL's properties basically consist of object property and data property. Meanwhile, annotation property is another type. However, object property clearly defines the relations between classes. As such, the property has to be diligently explored as it forms the basis on how intelligent an ontology based system is designed.

According to Neenu *et al* (2013); changes in temperature that can affect air vapor pressure deficits were identified, thus impacted the water use in agricultural landscapes. Similarly, the research of Kidane, (2018) also presented that plant having C4 photosynthesis reacts little to the rising of atmospheric CO_2 compare to that of C3 plant. This is due to the mechanism that increases the concentration of CO_2 in the leave which leads to CO_2 saturation of photosynthesis of this plant. However, in this paper more climatic conditions are considered. Some typical example of C4 crops are maize (corn), millet, sorghum, sugarcane and the like.

Considering Kuashik and Chatterjee (2017), the researchers proposed a schema for designing an ontology for agriculture domain with a focus on Indian context, the proposed schema worked in two step. In step one it uses domain-dependent regular expression and natural language procession techniques for automatic extraction of vocabulary related to the agricultural domain. Then the second step entails semantic relationships between the extracted terms and the candidate terms to be identified, which were carried out based on the rule-based reasoning algorithm called RelExOnt.

Moreover, Narayana, (2017) proposed a collaborative system that can support the management and sharing of multifaceted and large scale data source, which provide valuable and indispensable information for researcher. The proposed solutions relay semantic interoperability, construction of complex knowledge representation model bydesigning an ontology based decision support system for cotton crop farmers.

Furthermore, Bonacin *et al, (*2015) proposed an ontological model on impact of agriculture and climatic changes on water resource. This work describes interoperability issues in the engineering process of the OntoAgroHidro water resources. The research work emphasizes on the reuse of existing ontological models (Cauhsi and Sweet). However, in the case of this proposed paper, we created our ontological model for climatic conditions from scratch. A domain specialist read and analyzed each of the documents and marked the result as relevant or irrelevant. The paper presents representative scenarios and questions, and discusses the reuse and integration of concepts using knowledge visualization techniques. Experiments on the information recovery scenario point out the potential and limitations of the OntoAgroHidro.

Technically and based on literature, the application of ontology in agriculture is gradually gaining visibility like that of biomedicine. The research work of Malik et. al. (2018) developed fertilizer based ontology for the domain of agriculture and as well designed a user interface for query. The project is aimed to function in future as real time system by intending to merge it with existing ontologies within the domain such as crop or soil. Similarly, Chougule, et. al. (2019) advanced a research to developed an OWL based ontology recommended system for appropriate fertilizer to crop and appropriate soil for crop by considering a particular location in India.

Considering literature; there is no acceptable techniques for ontology validation and evaluation, which is very significant in ontology design as that determine the quality of ontology. Preference of technique is dependent on agreeing factors by ontology developers and/or domain experts (Walisadeera, et. al. 2016). Similarly, Dnyanesh and Rahul (2011) asserted in their work that the existing methodologies for ontology development fall short of comprehensive coverage for different activities such as pre-development, development and post-development stages. Consequently, to determine the appropriate methodology, some researches combined the activities of Methontology and Gruninger and Fox to design a given ontology (Iqbal, *et al.* 2013). Therefore, this literature forms the basis or the rationale behind the hybridization of the engineering processes used in this paper having reviewed the strengths and weaknesses of the existing work.

3.0 METHODOLOGY: THE ONTOLOGY ENGINEERING PROCESS

As earlier stated in the previous sections, there are available and different ontological development methodologies but not without shortcomings (Aminu et al. 2019). For example, while Methontology lacks competency question process, Gruninger-Fox lacks the process of ontology evolution and FAO-Based approach is limited in terms of specification of term or concepts. In view of these aforementioned limitations, the paper hybridized the three methodologies or ontology engineering processes to model the ontology under the domain of climatic conditions for maize crop farming. The hybridized approach used in this work consists of six processes namely as; collections of domain knowledge, specification of concepts (in this case, according to OWL's axioms), set competency questions to determine ontology's purpose and scope, ontology formalization, ontology validation and evaluation and ontology evolution as represented by Figure 1.

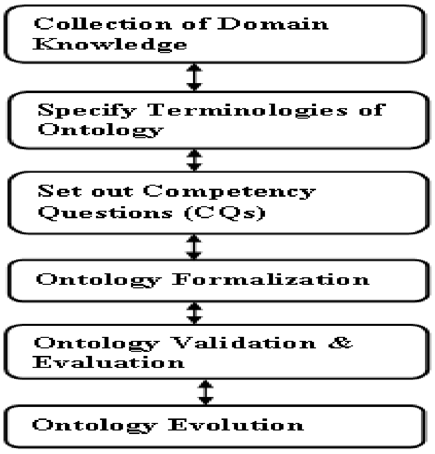

Collection of Domain Knowledge
Specify Terminologies of Ontology
Set out Competency Questions (CQs)
Ontology Formalization
Ontology Validation & Evaluation
Ontology Evolution

Figure1 : Ontology Engineering Process

For instance, the terms *maize, climaticCondition, Precipitation, Humidity, Sunlight, Rainfall, Wind* are specified classes components or sub concepts of the default class owl:Thing. Terms such as *affect* along with its synonyms *influences, change, involves* are specified and encoded as object property. Table1 presents some of the competency questions that would define the intelligent capability of the ontology as formulated and duly validated by both ontology developers and domain experts.

Table1: Ontology's Competency Questions

Competency question (CQs)	Contextual information required
What are the factors affecting maize?	• Is it the climatic condition during the pre-planting season? • Is it the climatic condition during the post-plant season? • Is it the climatic condition during the post-planting season
What are the factors that trigger physiological process and morphogenesis process?	• Physical attribute or characteristic involve. • Change occur based of color, height and the light • Simplex of complex development of maize
What are the factors that affect development process of maize?	• Stages of maize could be involving • Is it Physical development or biological development

Because of the model involved, it is expected that each query or question is put in context (meaning) as shown by the second column of the Table in order to achieve relevant precision and recall. In this section also we present the conceptual design or framework of the proposed system as shown by Figure2.

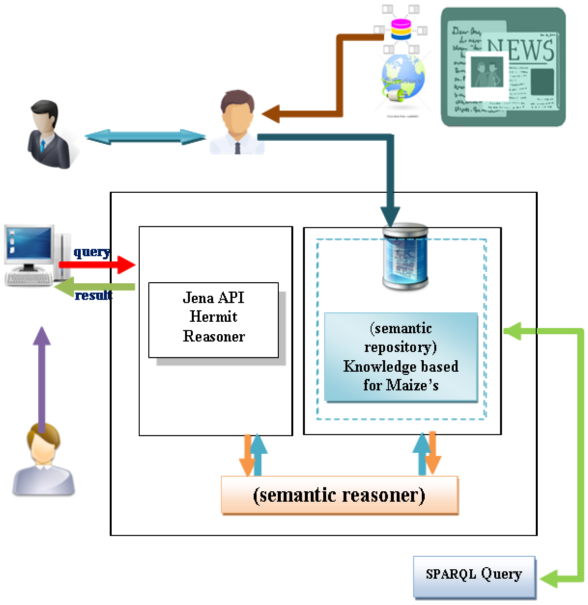


Figure 2: System Architecture of the Proposed

From the Figure 2, developer(s) collect domain related knowledge from various sources (such as trusted web documents, articles, books and institutions) and validated by domain experts. As soon as the knowledge are validated, the developer formalize the knowledge ontologically with the aid of Protégé editor and as well validated by its reasoners (Hermit, Pellet and ELK: reasoner) to develop the semantic repository. While developer uses SPARQL to query the ontology, a java based user interface is created taken into account the JENA Application Programming Interface (API) for end user's accessibility of the system. Besides, the proposed system's scenarios are furthered depicted by use case diagram of Figure3 and the data flow diagram of Figure 4.

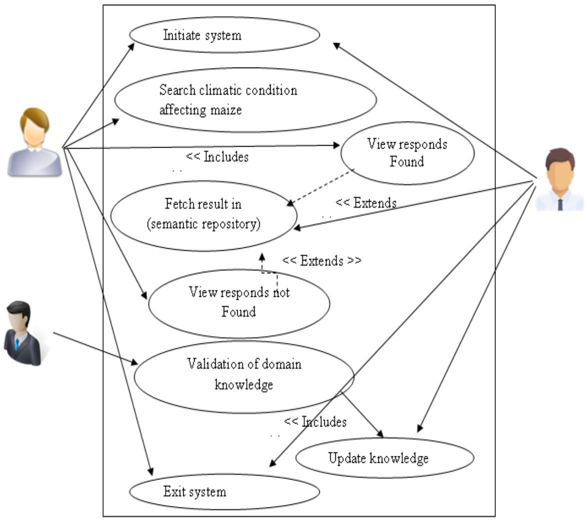


Figure 3: Use Case Diagram of the Proposed

Scenarios such as initiate or startup the system and search for climatic condition affecting maize can be performed by end user. While on one hand, ontology developer can engage in updating knowledge, domain experts on the other hand verify and validate the knowledge before being used by developer.

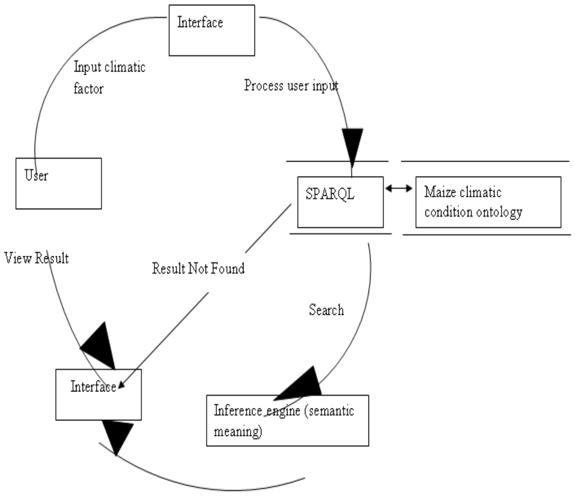


Figure 4: Dataflow Diagram of the proposed system

Considering the figure that describes how data serve as input or output for the next entity in the system. User input knowledge related to maize climatic condition through the interface as the front end. The communication between the front end and the ontology (back end) becomes possible as a result of the JENA API. Then the user interface forward user input for further processing by the reasoned. Ontology developer uses the SPARQL to perform query where the result can be displayed either at the back end or front end.

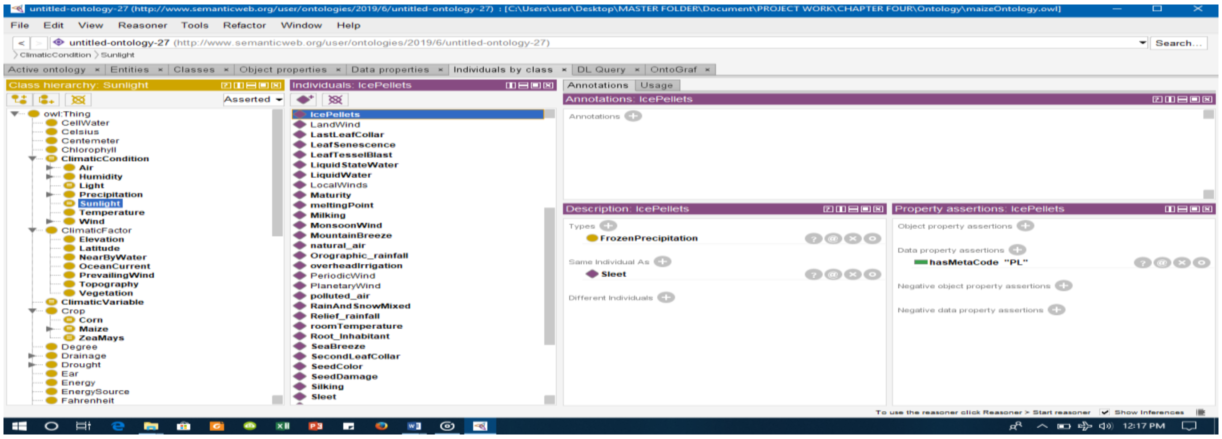
4.0 RESULTS AND DISCUSSION

This section accounts for the results of the ontology model of climatic condition for maize crop farming developed based on the axioms of OWL. As shown by Figure 5, terms have been specified as classes (both super and sub classes), relations (object property) and instances (individuals) appropriately and encoded using Protégé 5.2.2 version.

Figure 5 showed the graphical representation of the axioms specifically, the classes and how they are been related to each other via an object property with the aid of OntoGraf tab of protégé. The yellow colour circles represent the classes along with their various instances represent in purples. From the results of experiment currently, the metrics for the proposed ontology in general consists of 1150 axioms, 155 classes include the root class of owl:Thing. Climatic Conditions have 11 classes with 45 individuals and overall usage of 1526. This includes *climaticCondition* as a class and its other related or sub classes for instance, *rainfall, precipitation, wind, temperature and humidity*. Maize_Crop have 4 classes with 26 instances and a total usage of 2458 that is, the overall usage of the class *maize_crop* and other related classes or components. 119 object properties include owl:topObjectProperty the root property and 23 data properties include owl:topDataProperty. Although, the work is still ongoing, however, we were able to formulate fifteen numbers of competency questions which were validated by domain experts and users to evaluate the efficiency and effectiveness of the proposed system.

Figure 6 depicts the OntoGraph (Ontology Graph representation) represent a class using rectangular boxes and use arrow head line to indicate the relationship that exist between concept or class, in either symmetric manner or asymmetric. More so, with the aid of JENA API encoded in Java as shown in Figure7, we were able to develop a graphical user interface which made it easier for end users to use the ontology model to query or search related climatic conditions that can affect maize crop. The results of this feat are equally represented by Figure 8a and b.

Figure 8a represents the ontology user interface as search box which tested at the runtime of the system's execution. Figure8b had shown the results of user input query as "climatic condition affecting maize".



untitled-ontology-27 (http://www.semanticweb.org/user/ontologies/2019/6/untitled-ontology-27) : [C:\Users\user\Desktop\MASTER FOLDER\Document\PROJECT WORK\CHAPTER FOUR\Ontology\maizeOntology.owl]

ClimaticCondition 〉 Sunlight

Class hierarchy: Sunlight

- owl:Thing
  - CellWater
  - Celsius
  - Centemeter
  - Chlorophyll
  - ClimaticCondition
    - Air
    - Humidity
    - Light
    - Precipitation
    - Sunlight
    - Temperature
    - Wind
  - ClimaticFactor
    - Elevation
    - Latitude
    - NearByWater
    - OceanCurrent
    - PrevailingWind
    - Topography
    - Vegetation
  - ClimaticVariable
  - Crop
    - Corn
    - Maize
    - ZeaMays
  - Degree
  - Drainage
  - Drought
  - Ear
  - Energy
  - EnergySource
  - Fahrenheit

Individuals: IcePellets

IcePellets, LandWind, LastLeafCollar, LeafSenescence, LeafTesselBlast, LiquidStateWater, LiquidWater, LocalWinds, Maturity, meltingPoint, Milking, MonsoonWind, MountainBreeze, natural_air, Orographic_rainfall, overheadIrrigation, PeriodicWind, PlanetaryWind, polluted_air, RainAndSnowMixed, Relief_rainfall, roomTemperature, Root_Inhabitant, SeaBreeze, SecondLeafCollar, SeedColor, SeedDamage, Silking, Sleet

Annotations: IcePellets

Description: IcePellets

Types: FrozenPrecipitation

Same Individual As: Sleet

Property assertions: IcePellets

Data property assertions: hasMetaCode "PL"

Figure 5: OWL Axioms using Protégé

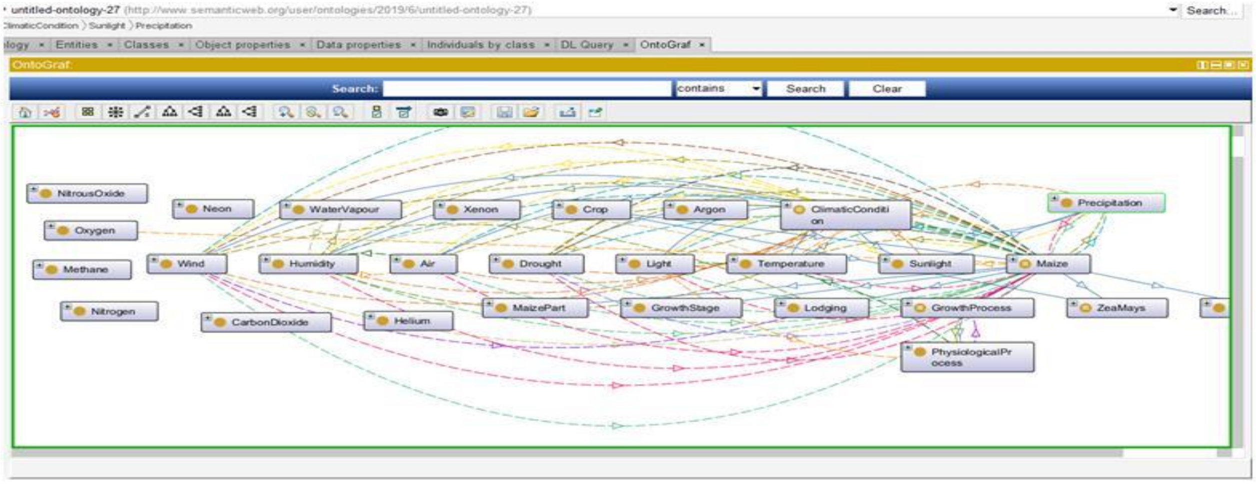
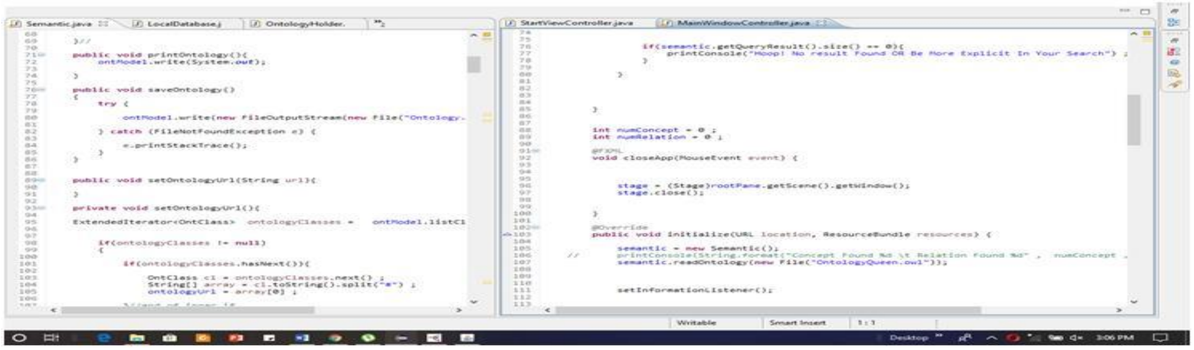


Figure 6: The Graph Representation of the Ontology



Semantic.java | LocalDatabase.j | OntologyHolder.

```
    }//

    public void printOntology(){
        ontModel.write(System.out);

    }

    public void saveOntology()
    {
        try {

            ontModel.write(new FileOutputStream(new File("Ontology.

        } catch (FileNotFoundException e) {

            e.printStackTrace();
        }
    }

    public void setOntologyUrl(String url){

    }

    private void setOntologyUrl(){

    ExtendedIterator<OntClass>  ontologyClasses =   ontModel.listCl

        if(ontologyClasses != null)
        {

            if(ontologyClasses.hasNext()){

                OntClass cl = ontologyClasses.next() ;
                String[] array = cl.toString().split("#") ;
                ontologyUrl = array[0] ;
```

StartViewController.java | MainWindowController.java

```
                if(semantic.getQueryResult().size() == 0){
                    printConsole("Hoop! No result Found OR Be More Explicit In Your Search") ;
                }

            }



        }

        int numConcept = 0 ;
        int numRelation = 0 ;

        @FXML
        void closeApp(MouseEvent event) {


            stage = (Stage)rootPane.getScene().getWindow();
            stage.close();


        }

        @Override
        public void initialize(URL location, ResourceBundle resources) {

            semantic = new Semantic();
    //      printConsole(String.format("Concept Found %d \t Relation Found %d" , numConcept ,
            semantic.readOntology(new File("OntologyQueen.owl"));


            setInformationListener();
```

Figure7: Java Based Code Fragment of the JENA

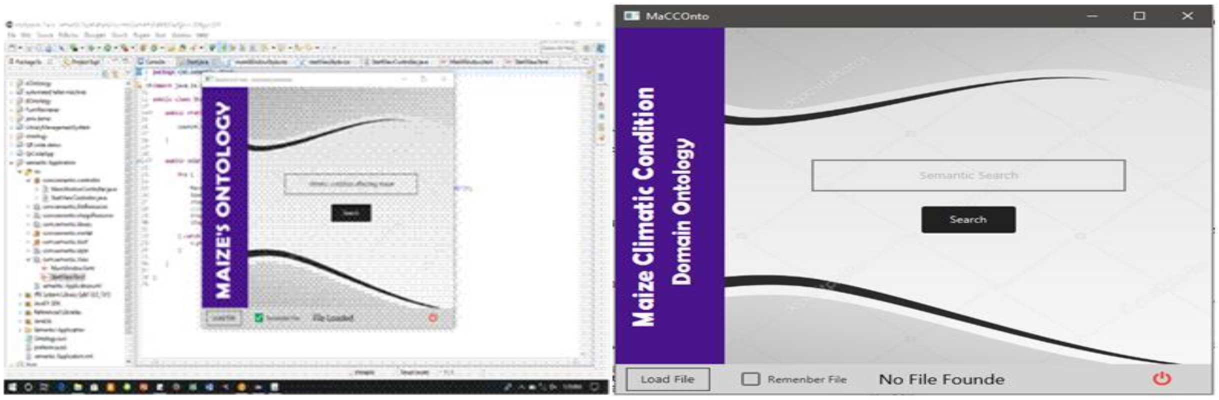
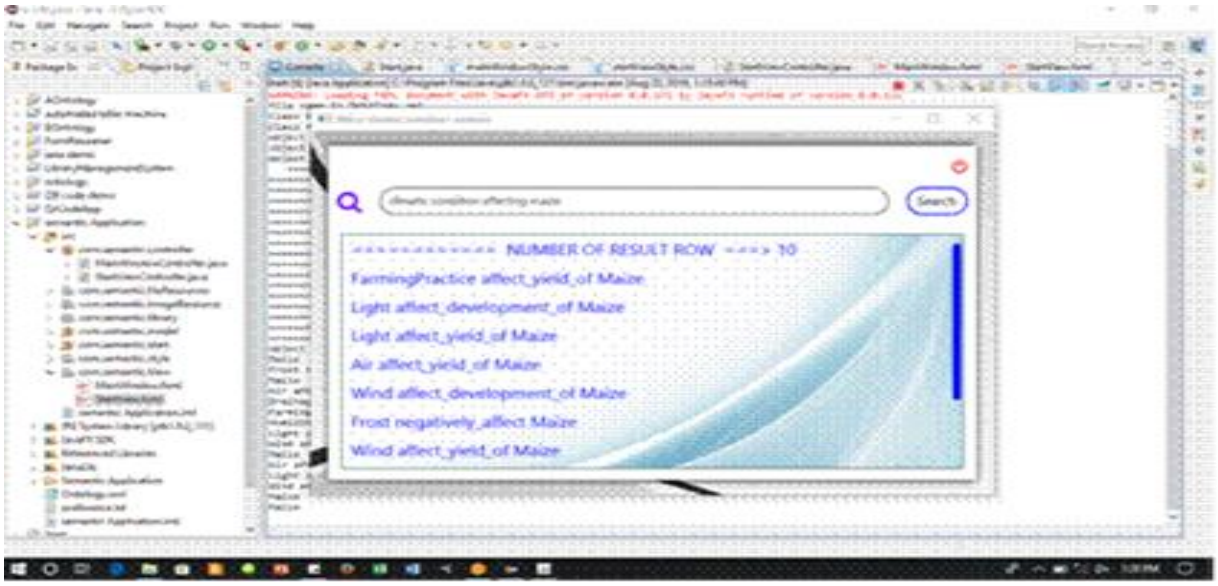


Figure 8a: Ontology GUI at Runtime



climatic condition affecting maize

========== NUMBER OF RESULT ROW ===> 10

FarmingPractice affect_yield_of Maize

Light affect_development_of Maize

Light affect_yield_of Maize

Air affect_yield_of Maize

Wind affect_development_of Maize

Frost negatively_affect Maize

Wind affect_yield_of Maize

Figure 8b: Ontology GUI at Runtime

5.0 SUMMARY AND CONCLUSION

In this research work, we model climatic conditions knowledge for maize crop farming. The maize farming ontology is developed using hybridization of Fox-Gruninger, Methontology and FAO-Based methodologies. The hybridization of these methodologies led to a six iterative ontology engineering process to model the ontology. It is created manually and written using OWL2 Web Ontology Language RDF/XML syntax. In the quest to design application based ontology in this project, the domain ontology is progressively developed into task ontology capable of answering questions in natural languages form. This research work aims to take ontology modeling beyond mere classification of domain knowledge like taxonomy but proposed to make the ontology an application support system; in order words, an intelligent system. This is so because, synonyms of object properties encoded in the ontology model were taken into account. In the overall experiment, the metrics for the ontology consists of 155 and 1150 classes and axioms respectively.

The competency questions were formulated and validated by the domain experts and were modelled into ontology design by considering about fifteen queries. At this point of the work, it has been validated based on experiments as indicated by the results and domain experts. The project is still work in progress because in the closest future, we hope to carry out evaluations on the efficiency of information retrieval considering precision and recall metrics. We will consider the metrics to judge the proposed project against the existing works. It is also recommended that an interested researcher can reuse and adapt the ontology. However, it is important to mention that this work only consider maize crop's climatic condition during planting season. This implies that further research on pre-planting and post-planting seasons can be carried out as well. That is, model climatic conditions knowledge for maize seeds and maize grains for pre-planting and post-planting seasons respectively.

REFERENCES

Alatrish, E. S. (2013). Comparison some of ontology. *Journal of Management Information Systems*, *8*(2), 018-024.

Aminu, E. F., Oyefolahan, I. O, Abdullahi, M. B., Salaudeen, M. T. (2019). A Review on Ontology Development Methodologies for Developing Ontological Knowledge Representation Systems for various Domains. IJEEB (Unpublished).

Antoniou, G., & Van Harmelen, F. (2004). *A semantic web primer*. MIT press.

Bonacin, R., Nabuco, O. F., & Junior, I. P. (2016). Ontology models of the impacts of agriculture and climate changes on water resources: Scenarios on interoperability and information recovery. *Future Generation Computer Systems*, *54*, 423-434.

Cardoso, J., & Pinto, A. M. (2015). The Web Ontology Language (OWL) and its Applications. In *Encyclopedia of Information Science and Technology, Third Edition* (pp. 7662-7673). IGI Global.

Choudhury, N. (2014). World wide web and its journey from web 1.0 to web 4.0. *International Journal of Computer Science and Information Technologies*, *5*(6), 8096-8100.

Chougule, A., Jha, V. K., & Mukhopadhyay, D. (2019). Crop Suitability and Fertilizers Recommendation Using Data Mining Techniques. In Progress in Advanced Computing and Intelligent Engineering (pp. 205-213). Springer, Singapore.

Da Silva, D. L., Souza, R. R., & Almeida, M. B. (2012). Ontologies and Controlled Vocabulary.

Dnyanesh R. & Rahul, C. (2011) A generic ontology development framework for data integration and decision support in a distributed environment, International Journal of Computer Integrated Manufacturing, 24:2, 154-170, DOI: 10.1080/0951192X.2010.531291.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

Iqbal R., Mustapha A., and Mohd. Y. Z., (2013),"An experience of developing Quran ontology with contextual information support", Multicultural Education & Technology Journal, Vol. 7 Iss 4 pp. 333-343.

Kaushik, N. & Chatterjee, N. (2018). Automatic relationship extraction from agricultural text for ontology construction. *Information processing in agriculture*, *5*(1), 60-73.

Kidane, B. (2018). Possible Consequences of Atmospheric CO_2 Level Increment on Chemical Composition and Functional Properties of Maize (Zea Mays L.) and Wheat (Triticum Aestivum L.) Varieties in the Past Thirty Years Collected from Ambo District, Amaro Kebele, West Shoa Zone of Oromia, Ethiopia (Doctoral dissertation, Addis Ababa University).

Malik, N., Sharan, A., & Shrivastav, J. (2018). Natural Language Interface for Ontology in Agriculture Domain. In Smart Trends in Systems, Security and Sustainability (pp. 259-268). Springer, Singapore.

Narayana, S. (2017). Ontology-based Advisory System for Cotton Crop Farmers. *International Journal of Computer Applications*, *167*(7), 1-7.

Neenu, S., Biswas, A. K., & Rao, A. S. (2013). Impact of climatic factors on crop production–A review. *Agric. Rev*, *34*(2), 97-106.

Sánchez, D., Isern, D., & Millan, M. (2011). Content annotation for the semantic web: an automatic web-based approach. *Knowledge and Information Systems*, *27*(3), 393-418.

Slimani, T. (2015). Ontology development: a comparing study on tools, languages and formalisms. *Indian Journal of Science and Technology*, *8*(24), 1-12.

Sokoto, M. B., Tanko, L., Abubakar, L., Dikko, A. U., Abdullahi, Y. M., & Unit, Z. (2016). Effect of climate variables on major cereal crops production in Sokoto State, Nigeria. *American Journal of Experimental Agriculture*, *10*(3), 1-7.

Walisadeera, A. I., Ginige, A., & Wikramanayake, G. N. (2016, July). Ontology evaluation approaches: a case study from agriculture domain. In *International Conference on Computational Science and Its Applications* (pp. 318-333). Springer, Cham.

Zacek, M. (2017, July). Ontology or formal ontology. In *AIP Conference Proceedings* (Vol. 1863, No. 1, p. 070012). AIP Publishing.

IMPACT OF BIOTECHNOLOGY AND INTEROPERABILITY IN HEALTHCARE INFORMATION SYSTEM

[1]Adedara R.O. and [2]Ajisola K.T.
Department of Computer Science, College of Education, Ikere, Ekiti State, Nigeia
[1]adedaraoluwadare@gmail.com, [2]asola_kolawole@yahoo.com

ABSTRACT

The growing health disparities between the recessed and the wealthy economy world call for urgent action from the scientific community. Science and technology have in the past played a vital role in improving public health. Today, with the tremendous potential of genomics and other advances in the life sciences, the contribution of science to improve public health and reduce global health disparities is more pertinent than ever before. Yet the benefits of modern medicine still have not reached millions of people in recessed economy countries. It is crucial to recognize that science and technology can be used very effectively in partnership with public health practices in recessed economy countries and can enhance their effectiveness. The fight to improve global health needs, in addition to effective public health measures, requires rapid and efficient diagnostic tools; new vaccines and drugs, efficient delivery methods and interoperability (Iop) of healthcare information system (HIS). Motivated by achieving a sustainable world, this paper discusses various technologies and issues regarding biotechnology and interoperability of health management in recessed economy.

Keywords: biotechnology, recessed economy, green health, interoperability.

1.0 INTRODUCTION

In the face of growing global health disparities, the important contribution of science and technology to improving health cannot be overlooked. Over the last 100 years, innovations in science and technology have resulted in improved health, quality of life, and a rise in life expectancy worldwide. Yet the benefits of modern medicine have still not reached millions of people in economy recessed countries. Biotechnology has tremendous potential to address health and development issues in economy recessed countries if we rise to the challenge (Tara, et al 2008). Contrary to its name, biotechnology is not a single technology. Rather it is a group of technologies that share two common characteristics (i) working with living cells and their molecules and (ii) having a wide range of practice uses that can improve our lives.

Chandrashekara et al (2015). Biotechnology can be broadly defined as use of cells or components of cells to produce products or processes for commercial purpose. As such, traditional biotechnology has been practices since the beginning of records history. It has been used to: bake bread, brew alcoholic beverages and breed food crops or domestic animals. But recent developments in molecular biology have given biotechnology new meaning, new importance and new potential. It is modern biotechnology that has captured the attention of the public, modern biotechnology can have a dramatic effect on the world economy and society.

Kevin et al (2015). One example of modern biotechnology is genetic engineering. Biotechnology began in the 1970s after the development of genetic engineering that allowed scientists to modify the genetic material of living cells. Genetic engineering is the process of transferring individual genes between organisms or modifying the genes in an organism to remove or add a desired trait or characteristic. Scientifically, genetic engineering is the manipulation of Deoxyribonucleic acids (DNA) molecules to produce modified plants, animals, or other organisms. DNA is the part of a cell that controls the genetic information of an animal or plant. DNA is a double-stranded molecule that is present in every cell of an organism. The genetic information is contained in individual units or sections of DNA called genes. The genes that are passed from parent to offspring determine the traits that the offspring will have. Scientists are now able to isolate the gene or genes for the traits they want in one animal or plant and move them into another. The movement of a gene from one organism to another is called recombinant DNA (rDNA) technology. It is this specific type of modern biotechnology (i.e., genetic engineering), that seems to generate the most attention and concern by consumers and consumer groups. What is interesting is that modern biotechnology is far more precise than traditional forms of biotechnology and so is viewed by some as being far safer.

2.0 RELATED WORK AND RESEARCH MOTIVATION

2.1 Applications of biotechnology in Healthcare

Though genetically modified food and crops have gained much international attention but modern biotechnology has been focusing on health sector for decades. Research and development in biotechnology applications of gene therapy both are related to health these days. Drug development and medical treatment are using most of the techniques of biotechnology. For example, both product and process have the basis of genetic engineering. Gene therapy is one tool which is used in the development of new product such as pharmaceuticals.

Biotechnical methods are now used to produce many proteins for pharmaceutical and other specialized purposes. As these genetically modified (GM) bacterial cells age, they produce human insulin, which can be purified and used to treat diabetes in humans. Microorganisms can also be modified to produce digestive enzymes. In the future, these microorganisms could be colonized in the intestinal tract of persons with digestive enzyme insufficiencies. Products of modern biotechnology include artificial blood vessels from collagen tubes coated with a layer of the anticoagulant heparin.

The following are the new, emerging biotechnologies and explore how they can be used to support the goals of economy recessed countries in improving health as well as other development indicators. Relevant examples would be used to identify ways in which these technologies can be used to achieve specific goals for health and development in economy recessed countries.

2.2 Recombinant DNA (rDNA)

Recombinant DNA (rDNA) technology has made a wide ranging influence in the area of human healthcare by legalizing the mass production of safe, pure and effective rDNA expression products. Currently, several categories of rDNA products i.e. hormones (for therapeutic use); haemopoietic growth factors; blood coagulation products; thrombolytic agents; antico-agulants and therapeutic enzymes are being produced using rDNA technology for human use. (Tara et al, 2008). Analytical methods play a vital role in the determination or confirmation of identity, purity and potency of rDNA products with respect to safe and efficacious medicine for human use. A summary of the analytical methods used in analysis of products is illustrated in Fig 1.

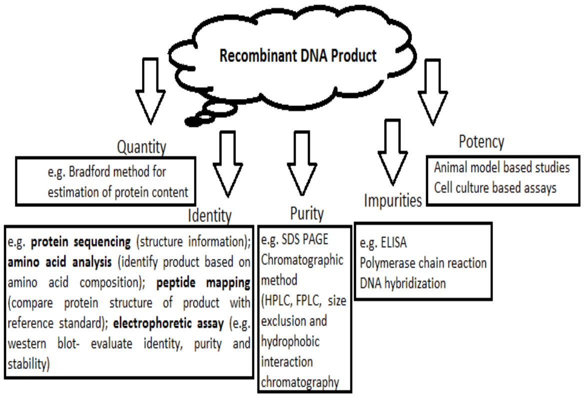

Recombinant DNA Product
Quantity
e.g. Bradford method for estimation of protein content
Identity
e.g. **protein sequencing** (structure information); **amino acid analysis** (identify product based on amino acid composition); **peptide mapping** (compare protein structure of product with reference standard); **electrophoretic assay** (e.g. western blot- evaluate identity, purity and stability)
Purity
e.g. SDS PAGE Chromatographic method (HPLC, FPLC, size exclusion and hydrophobic interaction chromatography
Impurities
e.g. ELISA
Polymerase chain reaction
DNA hybridization
Potency
Animal model based studies
Cell culture based assays

Figure 1: Methods used for the analysis of recombinant DNA products
(Source: Amit Gupta, Sushama R Chaphalkar (2016))

2.3 Therapeutics

Biotechnology helps in the development of treatments to cure diseases in mainly two ways: gene therapy and pharmacogenomics.

(a) Gene Therapy

Gene therapy can treat diseases by changing the genetic information of cells. Even though, gene therapy is new aspect, cases are there when patients have been cured by this technique.

Gene therapy mostly follows one of these approaches:

(i) the substitution of a mutated gene with a normal gene
(ii) the knocking out the activation of the mutated gene
(iii) Insertion of a completely new gene into the cells.

(b) Pharmacogenomics

Pharmacogenomics would help the doctors to recommend the correct medication and dosage on the basis of patient's genetic profile, thereby minimizing the risk of undesirable reactions, over dosage and side effects. Identification of SNPs is the basis of pharmacogenomics. Earlier, the sequencing of a genome was an expensive and lengthy procedure, but with the improvement of the DNA microarray, the sequencing has become easier and quicker. SNPs are being used to map and recognize specific genes that lead to the development of diseases such as cancer, arthritis and diabetes. The proteins produced by these genes become targets for novel therapies.

2.4 Paternity determination

This is possible because a child's DNA pattern is inherited, half from the mother and half from the father. To establish paternity, DNA fingerprints of the mother, child and the alleged father are compared. The matching sequences of the mother and the child are eliminated from the child's DNA fingerprint; what remains comes from the biological father. These segments are then compared for a match with the DNA fingerprint of the alleged father.

2.5 Diagnostics

For the past few years, a lot of progress in the field of animal and human disease diagnosis has been made and this has been brought by the advances in the area of diagnostic biotechnology i.e. diagnosis is now rapid, precise, less expensive and allows a broader range of epidemiological surveillance. In other words, biotechnology has played a key role in the identification of various infectious diseases that requires rapid, sensitive, specific and confirmatory recognition of pathogen.

The approaches that were commonly adopted included identification (causative agent isolated from pathogen) of samples collected after clinical observations and serological techniques (i.e. agglutination, precipitation, complement fixation and virus neutralization) in combination with the clinical observations. Detection of infectious diseases through recent advances areas of Biotechnology is illustrated in figure 2.

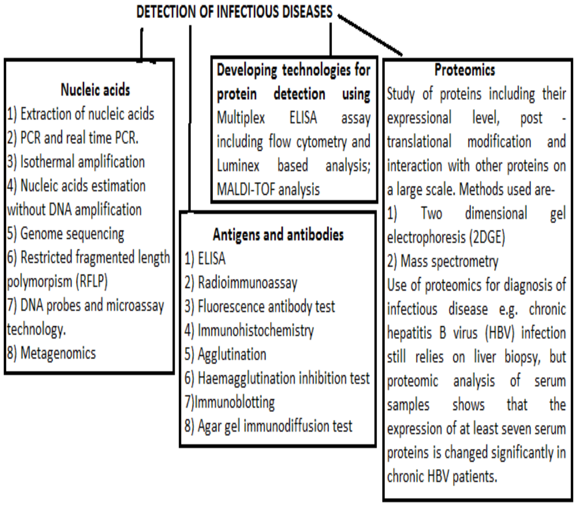

DETECTION OF INFECTIOUS DISEASES
Nucleic acids
1) Extraction of nucleic acids
2) PCR and real time PCR.
3) Isothermal amplification
4) Nucleic acids estimation without DNA amplification
5) Genome sequencing
6) Restricted fragmented length polymorpism (RFLP)
7) DNA probes and microassay technology.
8) Metagenomics
Developing technologies for protein detection using Multiplex ELISA assay including flow cytometry and Luminex based analysis; MALDI-TOF analysis
Antigens and antibodies
1) ELISA
2) Radioimmunoassay
3) Fluorescence antibody test
4) Immunohistochemistry
5) Agglutination
6) Haemagglutination inhibition test
7)Immunoblotting
8) Agar gel immunodiffusion test
Proteomics
Study of proteins including their expressional level, post - translational modification and interaction with other proteins on a large scale. Methods used are-
1) Two dimensional gel electrophoresis (2DGE)
2) Mass spectrometry
Use of proteomics for diagnosis of infectious disease e.g. chronic hepatitis B virus (HBV) infection still relies on liver biopsy, but proteomic analysis of serum samples shows that the expression of at least seven serum proteins is changed significantly in chronic HBV patients.

Fig.2. Detection of infectious diseases through recent advances areas of Biotechnology
(Source: Amit Gupta, Sushama R Chaphalkar

3.0 GREEN COMPUTING

Chandra et al (2016). Green computing or ICT, as per International Federation of Green (IFG) standard, is the study and practice of environmentally sustainable computing or IT. In, Murugesan defines the Green IT as “the study and practice of designing, manufacturing, using and disposing of computers, servers and associated sub systems such as monitors, printers, storage devices, networking and communications systems efficiently and effectively with minimal or no impact on the environment".

Green computing spans a number of focus areas and activities, including design for environmental sustainability, energy-efficient computing, power management, data center design, layout and location, server virtualization, responsible disposal and recycling, regulatory compliance, green metrics, assessment tools and methodology, environment-related risk mitigation, use of renewable energy sources and eco-labelling of its products.

3.1 Interoperability in Healthcare

Healthcare systems, in general, face unique challenges that include inadequate infrastructure, little or no access to continuous healthcare education, limited access to primary healthcare facilities and large distances from advanced care centers. Therefore, there is need for developing secure, efficient sharing of health information, not only for improving care, but also for engaging patients and supporting new models of care. Making progress on interoperability will support the aims of improving the patient experience of care (i.e., quality and satisfaction); improving the health of populations; and reducing the per capita cost of health care.

3.2 Identified Benefits of Interoperability

West Health Institute (WHI) and the Office of the National Coordinator for Health Information Technology (ONC) (2015) hosted Health Care Innovation Day (HCI) in Washington D.C. pointed out that interoperability is needed to reform the chaotic and at times dysfunctional nature of how information is shared among hospitals. Additionally, the report points out that unlike financial accounts and email, accessing your medical record cannot be done from virtually anywhere in the world. It's time for the healthcare system to catch up, but to do so in an organized way that prioritizes interoperability so that costly EMRs and other healthcare IT can communicate in a seamless way to the benefit of hospitals and patients alike.

Experience in healthcare indicates that interoperability can bring important benefits, including:

(i) Easier and Faster Access to Patients’ Information

With interoperable systems, data can be exchanged and stored automatically rather than re-typed into the system each time. This applies to all kinds of data used in healthcare: laboratory results, therapeutic procedures, medication administration, clinical notes, billing etc. This leads to: acceleration of communication, reduction in data recapture errors, reduction in duplicate efforts, etc.

(ii) Better diagnosis, better quality of treatment and better patient safety

Giving medical professionals faster access to patients’ data allows better diagnosis, better quality treatment, and better patient safety through: avoidance of medication interactions and better medication management to promote adherence and decrease abuse, improved knowledge of the patient health status, family history, personal history and better care coordination between the different healthcare professionals.

(iii) Improved cost efficiency

Interoperability between systems reduces administrative costs in manual data capture and reduction in workload for both clinical and administrative staff. Systems built on the same data exchange standards and using open access technologies are easier to integrate, reducing the implementation costs of new IT solutions in hospitals. Medical devices are able to transmit data directly to the Healthcare Information System (HIS) eliminates transcription errors and increase staff productivity.

(iv) More end to end security for data transfers

The exchange of patient data electronically requires privacy risks (identity theft, intrusion, alteration of data, and unauthorized access) to be addressed. Truly end to end interoperable IT systems with embedded privacy design reduce these risks through compatible security models, identification and authentication processes, data encryption etc.

3.3 System Design and Implementation

This section introduces the design methodology of the interoperability in Healthcare Information Systems (HIS)

4.0 DESIGN CONSIDERATIONS

(i) Interoperability: To pool different healthcare IT resources into large database so as to be able to share records easily.

(ii) Authentication: Security is achieved through the use of username and passwords.

4.1 Architecture of the System

The architecture of the interoperability for healthcare information system in figure 4. The system consists of three major components which are: models, views and controllers (MVC). This allows for each portion to be designed, implemented and tested independently from any other one, keeping code organized. Keeping the code organized means being able to find what is needed quickly, test features, correct or alter them quickly and add new functionality with ease. It also means more efficient code and a better way to re-use code for faster applications.

Probably one of the greatest benefits however is the consistency, managing a project between several developers can be easier as well.

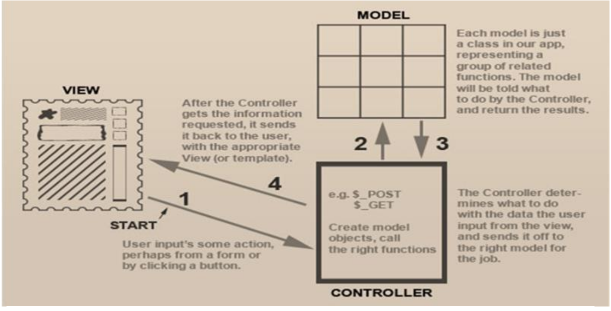

MODEL
Each model is just a class in our app, representing a group of related functions. The model will be told what to do by the Controller, and return the results.
VIEW
After the Controller gets the information requested, it sends it back to the user, with the appropriate View (or template).
2
3
4
1
START
User input's some action, perhaps from a form or by clicking a button.
e.g. $_POST $_GET
Create model objects, call the right functions
The Controller determines what to do with the data the user input from the view, and sends it off to the right model for the job.
CONTROLLER

Figure 3. Architecture interoperability for Healthcare Information System (HIS)

4.2 Implementation and Results

The implementation and testing of the system was carried out using five personal computers and one server meeting the minimum software and hardware requirements. The five computers represent five different users collaborating on the use of resources and sharing information. When the application is launched by the administrator, the administrator login page or authorization interface comes on. The hospital's administrator seeking authorization enters his username and password in their respective columns in order to gain access into the system; it compares the details with those of the administrator in the database and the system grants access to administrator, the login interface is shown in Figure 4.

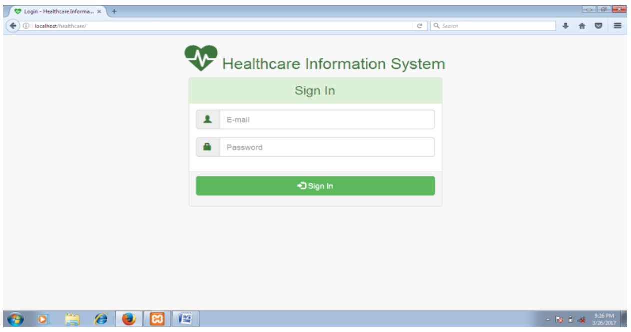
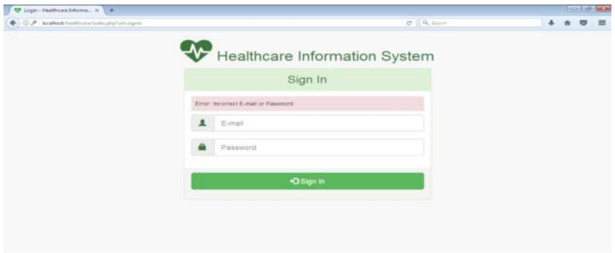


Figure 4. Login Interface for Administrator

If access is granted for an authorized user (i.e., administrator), it leads to another interface that displays the models or menus such as home, patients, medication, reports and management as shown in figure 4. An administrator has the privilege to register health actors (i.e., doctor, pharmacist, radiologist, clinicians etc) to the database and to view details (bio-data) of the health actors through the management menu.

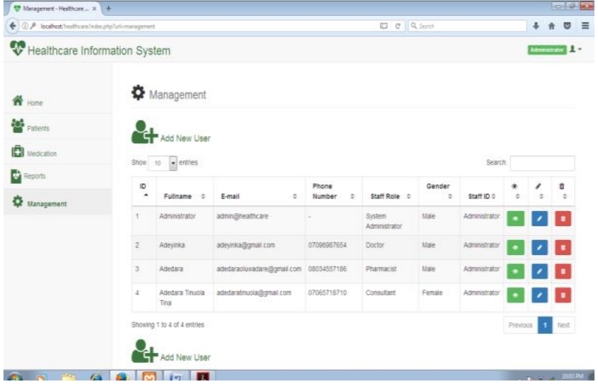
However, if an unauthorized user attempt to gain access into the system by entering incorrect details (i.e. incorrect username and password); the system denies such user the access as shown in Figure 5.



Error: Incorrect E-mail or Password

Figure 5. Unauthorized User Denied Access

When the administrator clicks on "management" menu, registered health actors profile such as doctor, pharmacist, radiologist clinicians etc (i.e., user) can be viewed and new users can be registered. This is shown in figure 6.



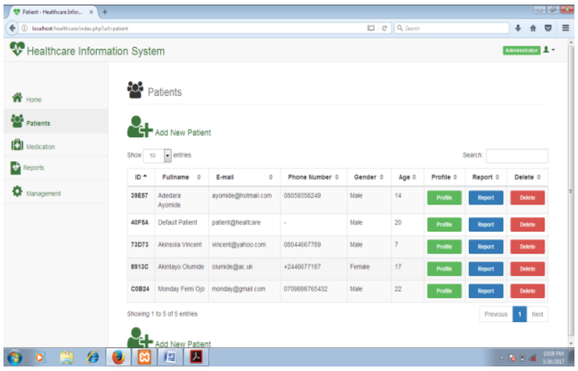
ID	Fullname	E-mail	Phone Number	Staff Role	Gender	Staff ID			
1	Administrator	admin@healthcare	-	System Administrator	Male	Administrator			
2	Adeyinka	adeyinka@gmail.com	07098987654	Doctor	Male	Administrator			
3	Adedara	adedaraoluwadare@gmail.com	08034557186	Pharmacist	Male	Administrator			
4	Adedara Tinuola Tina	adedaratinuola@gmail.com	07065718710	Consultant	Female	Administrator			

Showing 1 to 4 of 4 entries

Previous 1 Next

Figure 6. Interface for an Authorized

The administrator also has the right to register new patients or view the profile of existing patients, this is shown in figure 7.



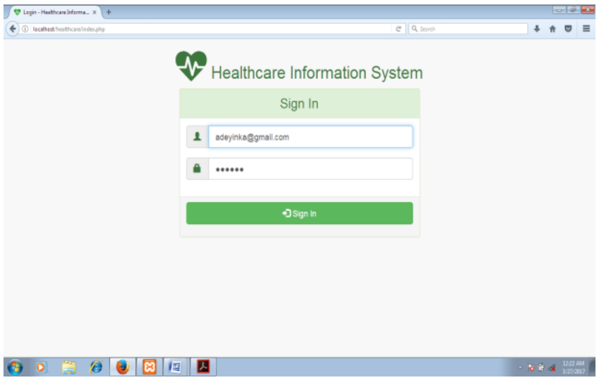
ID	Fullname	E-mail	Phone Number	Gender	Age	Profile	Report	Delete
39E57	Adedara Ayomide	ayomide@hotmail.com	08059358249	Male	14	Profile	Report	Delete
40F5A	Default Patient	patient@healtcare	-	Male	20	Profile	Report	Delete
73D73	Akinsola Vincent	vincent@yahoo.com	08044667789	Male	7	Profile	Report	Delete
8913C	Akintayo Olumide	olumide@ac.uk	+2446677187	Female	17	Profile	Report	Delete
C0B24	Monday Femi Ojo	monday@gmail.com	0709898765432	Male	22	Profile	Report	Delete

Showing 1 to 5 of 5 entries

Previous 1 Next

Figure 7. Interface showing the list of existing patients and add new patient

When the application is launched by the health actors (e.g., doctor or pharmacist), the doctor's login or authorization interface comes on. The doctor seeking authorization enters his username and password in their respective columns in order to gain access into the system; it compares the details with those of the doctor in the database and the system grants access to doctor, the login interface is shown in Figure 8.

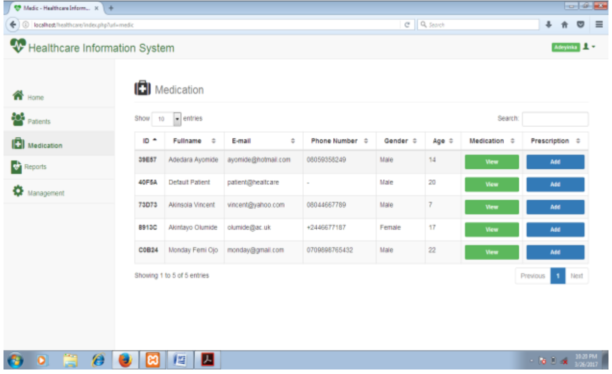

Login - Healthcare Informa...
localhost/healthcare/index.php
Healthcare Information System
Sign In
Sign In

adeyinka@gmail.com

••••••

Figure 8. Login Interface for Health Actors

The doctor can give medication through the "medication menu", the list of all patients will come out as shown in figure 9.



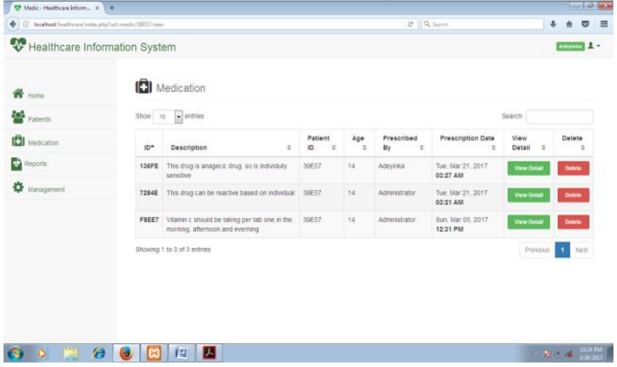
Show 10 entries Search:

ID	Fullname	E-mail	Phone Number	Gender	Age	Medication	Prescription
39E57	Adedara Ayomide	ayomide@hotmail.com	08059358249	Male	14	View	Add
40F5A	Default Patient	patient@healtcare	-	Male	20	View	Add
73D73	Akinsola Vincent	vincent@yahoo.com	08044667789	Male	7	View	Add
8913C	Akintayo Olumide	olumide@ac.uk	+2446677187	Female	17	View	Add
C0B24	Monday Femi Ojo	monday@gmail.com	0709898765432	Male	22	View	Add

Showing 1 to 5 of 5 entries

Figure 9. Patients' Medication Interface

From figure 9 above, when doctor clicks on "view" menu for a particular patient, the package display all the lists (i.e., health history) of medications and prescriptions giving to that patient as shown in figure 10.


Medic - Healthcare Inform...
localhost/healthcare/index.php?url=medic/39E57/view
Search
Healthcare Information System
Adeyinka
Home
Patients
Medication
Reports
Management
Medication
Previous
1
Next
10:24 PM
3/26/2017

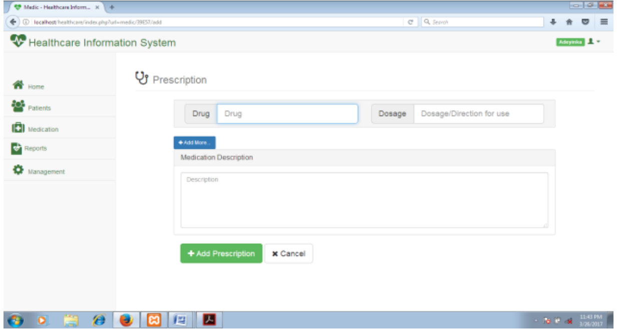
Show 10 entries Search:

ID	Description	Patient ID	Age	Prescribed By	Prescription Date	View Detail	Delete
136FE	This drug is anagecic drug, so is individuly sensitive	39E57	14	Adeyinka	Tue, Mar 21, 2017 **03:27 AM**	View Detail	Delete
7284E	This drug can be reactive based on individual	39E57	14	Administrator	Tue, Mar 21, 2017 **03:21 AM**	View Detail	Delete
F8EE7	Vitamin c should be taking per tab one in the morning, afternoon and evening	39E57	14	Administrator	Sun, Mar 05, 2017 **12:31 PM**	View Detail	Delete

Showing 1 to 3 of 3 entries

Figure 10. Patient' Medications details through "View detail" menu

From figure 10 above, when doctor clicks on "add menu" under prescription column, doctor can prescribe drug(s), give description of drugs' usage to patient as shown in figure 11.



Prescription

Drug: Drug

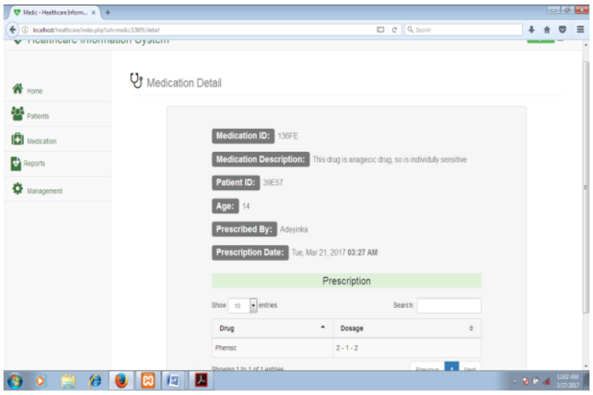
Dosage: Dosage/Direction for use

Medication Description

Description

Figure 11. "add prescription" interface

From figure 11 above, when doctor clicks on "view details" menu, it displays a report of patient's medication detail as shown in figure 12, the medication ID is written on paper by the doctor, given to the patients and present it to pharmacist.


Medic - Healthcare Inform...
localhost/healthcare/index.php?url=medic/136FE/detail
Search
Healthcare Information System
Home
Patients
Medication
Reports
Management
Previous
1
Next
12:02 AM
3/27/2017

Medication Detail

Medication ID: 136FE

Medication Description: This drug is anagecic drug, so is individuly sensitive

Patient ID: 39E57

Age: 14

Prescribed By: Adeyinka

Prescription Date: Tue, Mar 21, 2017 **03:27 AM**

Prescription

Show 10 entries Search:

Drug	Dosage
Phensic	2 - 1 - 2

Showing 1 to 1 of 1 entries

Figure 12. Medication Detail interface

5.0 CONCLUSION AND RECOMMENDATIONS

The applications of biotechnology are so broad, and the advantages so compelling, that virtually every industry is using this technology. Biotechnology is enabling these industries to make new or better products, often with greater speed, efficiency and flexibility. Biotechnology holds significant promise to the future but certain amount of risk is associated with any area. Biotechnology must continue to be carefully regulated so that the maximum benefits are received with the least risk.

The future of the world food supply depends upon how well scientists, government and the food industry are able to communicate with consumers about the benefits and safety of the technology.

Also, strong health systems are central to achieving better health outcomes and strong Healthcare Information Systems (HIS) are the backbone of strong health systems. A properly functioning HIS gets the right information into the right hands at the right time, enabling policymakers, managers and individual service providers to make informed choices about everything from patient care to national budgets. Significant care coordination gaps exist due to the lack of interoperability across the low and middle-income countries (LMICs). There is need to design, evaluate and incentivize the use of Healthcare Information Technology (HCIS) for effective care coordination.

REFERENCES

1. **Amit Gupta, Sushama R Chaphalkar (2016).** *Role of Biotechnology in Human Health Care.* Emer Life Sci Res, 2(1): Pp. 67-72.
2. **Chandra et al (2016).** Green IoT Agriculture and HealthcareApplication (GAHA). *International Journal of Smart Home*. Vol. 10, No. 4 (2016), pp. 289-300. http://dx.doi.org/10.14257/ijsh.2016.10.4.26
3. **Chandrashekara et al (2015).** Basic Concept of Biotechnology. LAXMI Book Publication. Solapur, Maharashtra, India. Pp 322-333.

4. **Tara et al. (2008).** *Biotechnology to improve health in developing countries - A review.* Joint Centre for Bioethics, University of Toronto, Canada. http://dx.doi.org/10.1590/S0074-02762004000400001
5. **Kevin et al (2015).** *Biotechnology and its Applications.* North Carolina State University**,** College of Agriculture and Life Science. Pp 1 - 11
6. **Office of the National Coordinator for Health Information Technology (ONC) (2015).** *Electronic Health Record Vendors Reported by Hospitals Participating in the CMS, EHR Incentive Programs,'* Health IT Quick-Stat.gov/quickstats/pages/FIG-Ven-dors-of-EHRs-to-Participating-Hospitals.php. June, 2015.

USER ACCEPTANCE OF INTERNET BANKING IN NIGERIA

[*]Aminu S. A., Alasiri W. A., Cole A. A. and Adeoti I. O.
[a]Department of Marketing, Lagos State Polytechnic, Ikorodu, Lagos State, Nigeria.
[*]Corresponding author's e-mail: asabiod2001@yahoo.com.

ABSTRACT

Since the Internet technology broke out in the global arena several years ago, it has transformed the ways firms in different industries conduct their businesses. The banking industry in Nigeria is leveraging on the unique benefits of the Internet to deliver banking services online and provide an alternative delivery channel to the traditional channel of distributing financial services. However, the population of bank customers using Internet banking is still small. The paper investigated the factors responsible for the acceptance of Internet banking in Nigeria, using an extended technology acceptance model (TAM). Using student data from a valid sample of 288 users of Internet banking from Lagos State Polytechnic and analysed with a multiple regression method, the results provided support for three of the four factors examined while the last factor was not supported by our dataset. The analysis suggested that perceived usefulness (PU), perceived enjoyment (PE) and website security & privacy have a significant effect on Internet banking acceptance. Surprisingly, perceived ease of use (PEOU) did not have a significant effect on Internet banking acceptance. This paper contributed to the existing knowledge on Internet banking acceptance in a developing country context like Nigeria. The paper highlighted the managerial implications of the findings and discussed their limitations.

Keywords: Internet banking acceptance, perceived usefulness, perceived ease of use and technology acceptance model.

1.0 INTRODUCTION

Following the breaking of technology in the global arena, its application and use has known no bounds and has permeated all industries in both the developed and developing countries. Technology has the greatest impact on business activities and is an important source of competitive advantage for many companies across different industries (Aminu, 2013). The Internet is a major technological breakthrough, transforming the ways goods and services are being produced and distributed in different industries. A recent study identified the Internet as a unique technological platform facilitating a direct interface, engagement and relationship between marketers and their customers (Aminu et al., 2018). The Internet has not only been a catalyst for the development of new industries, but it has also transformed the ways many other industries do business (Hernández-Murillo et al., 2010). The Internet will become a full-fledged delivery and distribution channel for many products, with electronic financial products and services leading the pack (Tan and Thompson, 2000). Truly, this prediction has come to pass, almost three decades after the prediction was made as the capability of the Internet to deliver banking and other financial services is almost limitless.

Like many industries around the world, the banking industry in Nigeria is leveraging on the unique benefits of the Internet to deliver financial services, in addition to the traditional banks' branches, where most customers in the country are currently accessing banking services. This is called Internet banking in the Internet banking literature. Internet banking and online banking are used interchangeably in the literature. Internet banking is an Internet portal, which customers can use for different kinds of banking services ranging from bill payment to making investment (Pikkarainen et al. 2004).

Since the arrival of the Internet in 1995, it has simplified banking and enables banks to offer new services to their customers, further making a visit to the branches unattractive (Aminu et al., 2018). Therefore, many banks, including those in Nigeria, have developed a high technological infrastructure and capability to offer financial services online. A study found that, in addition to banks and governments' supports, customer acceptance of Internet banking is an important determinant of the success of Internet banking (Sathye, 1999). Another study affirmed that the success of an information system in the market is determined by the degree to which the system is used and ultimately adapted by the users (Venkatesh et al., 2003). Therefore, a growing number of researchers are interested in a new technology adoption. It is noted that adoption or acceptance of a new product is an important topic in Marketing (Aminu, 2018).

Internet banking acceptance, especially in developing countries such as Nigeria, is still low. KPMG (2017) observed that despite a great deal of investment and efforts banks in Nigeria have made to achieve an effective online banking ecosystem and the existence of Internet savvy bank customers, converting these customers to users of Internet banking remains a problem. KPMG's survey suggested that more than two-thirds of Nigerian banking customers say they have never tried their banks' online platforms, suggesting that just one-in-three of them have ever conducted banking transactions online (KPMG, 2017). More than half of the people who have tried Internet banking services did not become regular users (Lallmahamood, 1970). Non-adopters have expressed concerns about ease of use of Internet banking as well as security of web-based channels (KPMG, 2017). The problem of slow web navigation (Forsythe and Shi, 2003; Ajuwon, 2015), often causing unpleasant experience, may be discouraging a large number of customers from adopting Internet banking. Therefore, banks' marketers need the knowledge of factors that drive the acceptance of Internet banking in Nigeria and emphasise these in their marketing communication strategy.

We examine the influences of four factors on acceptance of Internet banking in Nigeria using student data. These factors include the two widely used TAM's constructs - perceived usefulness (PU) and perceived ease of use (PEOU) (Davis, 1989) and additional two factors - perceived enjoyment (PE) and website security and privacy suggested in the model by Pikkarainen et al. (2004). Due to the few empirical studies on Internet banking acceptance in Nigeria and inconsistencies in the findings of these study, this paper contributes to the research on Internet banking acceptance in Nigeria and further provides new empirical findings on the Internet phenomenon in the country. The findings of this research will have far-reaching implications for the marketers of banks in Nigeria and other developing countries. The rest of the paper includes literature review, methods, results and conclusion.

2.0 LITERATURE REVIEW AND DEVELOPMENT OF HYPOTHESES

2.1 Internet Banking

Internet banking and online banking are used interchangeably in the Internet banking literature. This paper follows this tradition in the conceptualisation of Internet banking. Internet banking is an offer that allows customers to perform a wide range of banking transactions electronically via the bank’s website (Tan and Thompson, 2000). Internet banking service referred to a banking service that allows customers to access and perform financial transactions on their bank accounts from their web enabled computers with Internet connection to banks' websites any time they wish (Ongkasuwan Tantichattanon, 2002). Online banking is Internet portal which customers can use for different kinds of banking services ranging from bill payment to making investment (Pikkarainen et al., 2004). The online banking system is a part of the banking ecosystem operated by a bank, but is an alternative to branch banking through which customers commonly use banking services (Oke, 2019). Customers' interface with the bank websites allows them to open accounts, apply for loans, check balances, transfer funds and make and receive payments on the Internet (Salawu and Salawu, 2007). From the foregoing, Internet banking allows bank customers to conduct banking transactions in a virtual environment devoid of a physical visit to the banks' branches for the purpose of interacting with the banks' personnel.

Compared to other platforms of electronic banking (e-banking), Internet banking is widely perceived as the most important and widely used delivery channel for banking services in the Internet age (Cheng et al., 2006). The rapid penetration of the Internet among the users is increasingly promoting the banks' websites as an alternative delivery channel for banking products and services (Tan and Thompson, 2000). Banks adopt Internet banking to enhance the effectiveness of distribution channels through reducing the transaction cost and increasing the speed of services (Al-Smadi, 2012). Cost savings and widespread adoption of the Internet have promoted Internet banking development and usage (Pikkarainen *et al*., 2004; Hernández-Murillo et al., 2010). Internet banking requires less staff and fewer physical branches (Cheng *et al*., 2006). Therefore, Internet banking does not require the huge investment in buildings, facilities, equipment, supplies and personnel, associated with the traditional brick-and-mortal banking format. This has significantly contributed to the cost reduction associated with delivering banking services via the banks' websites, often cited in the literature.

From the customers' perspectives, Internet banking provides customers the benefits of convenience, speed, flexibility, freedom of location and 24-hour availability of banking services (Cheng et al., 2006; Gerrard, Cunningham and Devlin, 2006; KPMG, 2017). With the increasing access to mobile devices, banks' customers can quickly and easily conduct banking transactions from the comfort of their homes, cars, mass transit buses, offices, market places, shopping malls, classrooms and any other place they choose.

2.2 Internet banking in Nigeria

The Nigerian banking ecosystem is changing. It is being increasingly characterised by upsurge in the use of the Internet and other cutting edge technological innovations in the financial services sector (Ajuwon, 2015). Internet banking services were first introduced into the Nigerian financial system in 2001. Thereafter, other electronic banking services such as the ATM and phone banking were introduced (Agwu and Carter, 2014). The Central Bank of Nigeria (CBN), in 2002, found out that of the 89 licensed banks in the country, 17 were offering Internet banking, 24 were offering basic telephone banking, seven had ATM services, while 13 of the banks were offering other forms of e-banking. Therefore, only 19 percent of the banks were offering Internet banking (Ezeoha, 2005).

This statistics has improved considerably with the banking reform of mid-2000. Aminu (2012) explained that the banking consolidation, ending in December 2005, significantly enhanced the capital base of the successful banks and fiercely increased the competition among them. The recapitalisation enabled the banks to invest in infrastructure of all kinds (Aminu, 2012). Poised to raise the quality of banking to the global standards, banks in Nigeria invested massively on technology and widely adopted electronic and telecommunication networks for delivering an array of financial services (Chiemeke et al., 2006). The development of the Internet in Nigeria has provided customers with the means to carry out virtual business activities (Oni and Ayo, 2010). One of the greatest impacts of Internet banking in Nigeria is retail banking spikes and growth in the use of e-banking delivery channels (Ajuwon, 2015).

The revolution of Internet banking in Nigeria is considerably facilitated by the great leap in the number of Nigerians that have adopted the Internet. The latest Internet statistics estimated the population of Internet users in the country to be 116.3 million as at December, 2018 (Internet World Statistics, 2019). However, while millions of Nigerians are on the Internet chatting, only few of them, about one-third, are using their Internet for banking transactions (KPMP, 2015). Internet banking is yet to gain much traction in the country because of certain problems. These include the challenges of slow navigation, fraud, insecurity, high level of illiteracy, culture and trust deficit (Ezeoha, 2005; Chiemeke et al., 2006; Aminu, 2013; Oke, 2019). It is certain that the Internet banking revolution in the country will continue as these and other impediments to Internet banking are removed and with the continuous engagement of the bank customers by the banking and other financial institutions through marketing communications.

2.3 The Technology Acceptance Model and Internet Banking Acceptance

Internet banking acceptance, like many other technologies' acceptance, has been widely explained by a number of adoption theories. Some of these theories are the theory of reasoned action (TRA) (Fishbein and Ajzen, 1975), the theory of planned behaviour (TPB) (Ajzen, 1985), diffusion of innovation (DOI) theory (Rogers, 1995) and TAM (Davis 1989; Davis et al., 1989). All these theories are useful in the understanding of the factors influencing a technology usage. This present study is anchored on an extended model of TAM developed by (Pikkarainen et al., 2004). This model extends Davis' TAM.

The original TAM by Davis (1989) is based on two cognitive constructs - PEOU of an information system (IS) and PU of IS. According to Davis, these two constructs can be used to explain and predict attitude (ATT) of potential users towards a new IS, their behavioural intention (BI) and actual usage (AU). TAM is an extension of TRA (Fishbein and Ajzen, 1975). Davis explained that the user acceptance of a new IS is determined by PEOU and PU of the system. PEOU referred to a degree to which a user of a system believes that using the system will be free of effort while PU described the extent to which a user believes that a particular system would improve his/her job performance. These cognitive beliefs lead to attitude formation toward the system, a positive attitude creates BI for the system, which results in actual usage of the system (Davis, 1989). Empirical evidence from TAM indicated that it explain much of the variance in usage intention and self-reported usage (Davis 1989; Davis et al., 1989). This explains its popularity among the technology adoption researchers.

However, it is believed that PEOU and PU alone may not be sufficient to explain the user's intention to adopt electronic banking (Oni and Ago, 2010). Therefore, a number of authors and researchers have introduced additional constructs to the two cognitive constructs of PEOU and PU in the original model to enhance the explanatory power of the model. One of these authors is Pikkarainen et al. (2004). There few of these in Nigeria. This paper extends the few literature on Internet banking adoption.

In a tested model explaining the acceptance of Internet banking in Finland, Pikkarainen et al. (2004) extended the original TAM (Davis, 1989) and added new constructs - PE of Internet banking, security and privacy of Internet banking, Internet banking information and quality of Internet connection - to the model. We combined two of these constructs, perceived enjoyment and security and privacy of websites with PEOU and PU in this paper. These four constructs are used to operationalise the independent variables of this paper and are briefly reviewed below.

(a) Perceived usefulness (PU)

PU described the extent to which a user believes that a particular system would improve his/her job performance (Davis, 1989). Individuals would use an application if they believe that it would facilitate their performance (Oni and Ago, 2010). PU is considered to be more important than PEOU because users of a new technology would first consider its usefulness before getting to know how to use it and accepting its usage (Gardner and Amoroso, 2094). PU of Internet banking is one of the reasons why customers switch from traditional banking to Internet banking (Qureshi, Zafar and Khan, 2008). A study in Bangladesh showed that PU significantly explained the reason why banks' customers in the country adapted Internet banking (Jahangir and Begum, 2008). Another one in Jordan found that PU had a positive and significant effect on customer attitude towards electronic banking usage (Al-Smadi, 2012). Similarly, it was proven that PU influenced an individual's attitude towards using a technology (Nath et al., 2014). Pikkarainen et al. (2004) found that PU was one of the two main factors influencing Internet banking acceptance in Finland. Their results suggest that users use Internet banking for the tangible benefits it offers over other banking delivery channels. Finally, a study in Hong Kong revealed that PU had a direct effect on the behavioural intention of users to use Internet banking (Cheng et al., 2006). In line with this review, we postulate that:H1: PU has a significant effect on user acceptance of Internet banking.

(b) Perceived ease of use (PEOU)

PEOU is a degree to which a user of a system believes that using the system will be free of effort (Davis, 1989). The construct suggests that the adoption of a new technology will be facilitated by the ability of the user to quickly learn to use the technology with little or no training. It is believed that the users of point of sale (PoP) devices require no training to use them (Aminu, 2018) and this may be facilitating its increasing adoption in Nigeria. Similarly, learning to conduct banking transactions on the Internet does not require extensive training on the part of the users. According to Davis, PEOU influences the PU because a technology that is easier to use will become useful for the user (Davis, 1989). It is affirmed that adopters would use a new application if they believe it is easy to use (Oni and Ayo, 2010). PEOU influenced an individual's attitude towards technology usage (Nath et al., 2014) and is one of the reasons why customers adopt Internet banking (Qureshi et al., 2008). An empirical study in Hong Kong found that PEOU was a significant secondary determinant of customer intention to use Internet banking (Cheng et al., 2006). In Bangladesh, the research by Jahangir and Begum (2008) showed that PEOU was significantly and positively related to the reason why banks' customers in the country switched from the traditional banking to

Internet banking. The finding of another study in Jordan revealed that PEOU had a positive and significant effect on customer attitude to use electronic banking (Al-Smadi, 2012). PEOU had a significant effect on Internet banking use with smartphones (Yoon and Occeña, 2014). Conversely, a study showed that PEOU had an indirect effect on the intention of banks' customers to use Internet banking in Hong Kong (Cheng et al., 2006). Another indicated that PEOU did not significantly affect Internet banking acceptance in Finland (Pikkarainen et al., 2004). In view of the foregoing, we state that:

H2: PEOU has a significant effect on user acceptance of Internet banking.

(c) Perceived enjoyment (PE)

Though shopping is widely perceived as a major leisure-time activity, many shoppers complain about their shopping experiences. Shoppers, while engaging in shopping, consider overcoming boredom, peer group influence and status consciousness (Reid and Brown, 1996). Therefore, PE may be an important consideration for many consumers when they are making purchases in both the physical and virtual stores. According to Davis et al. (1992), PE referred to the extent to which the activity of using the computer is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated. Due to the extensive use of technology in the Internet transactions, compared to the traditional in-store transactions (Naiyi, 2004), Internet transactions increase the shoppers' experiences than in-store shopping (Forsythe and Shi, 2003). The attitude of a user of a new technology, who enjoys using the technology, is likely to be positive, resulting in the user repeating a pleasant activity and avoiding an activity that is less enjoyable (Suki and Suki, 2011). PE had a strong relationship with Internet usage in Korea (Teo et al., 1998). PE was one of the direct determinants of customers' attitude and behavioural intentions to use Internet banking in India. Further, the result suggested that PE exerted both direct and indirect effects on PU but exerted only direct effect on PEOU (Bashir and Madhavaiah, 2015). A recent empirical study also showed that PE strongly predicted behavioural intention, stronger than perceived ease of use (Hormbaek and Hertzum, 2017). In contrast, Davis (1989) found that PE did not significantly affect Internet banking acceptance in Finland. PE did not have a positive effect on the behavioural intentions of Malaysians towards using 3G mobile services (Suki and Suki, 2011). From the foregoing, we propose that:

H3: PE has a significant effect on user acceptance of Internet banking.

(d) Website security and privacy

Security is defined as the protection of information or systems from unauthorised intrusions (Egwali, 2008). Internet shoppers may feel that their private information is not guaranteed, giving rise to the perception of privacy risk. Featherman et al. (2010) defined privacy risk as a degree to which a consumer believed that his/her personal information may be lost to or misused by others, resulting in identity theft. Internet shoppers' personal information may be more vulnerable to theft and misuse than the traditional shoppers' personal information (Aminu et al., 2018). Concern for security among customers was a leading barrier for non-acceptance of Internet banking in Latin Americs (Booz Allen and Hamilton, 1997). Security on vendors' websites is often cited by users as a major factor that eliminates the barriers to shopping on the websites (Salisbury et al., 2001). Therefore, customers expect to be secured and comfortable with Internet banking because its operations are substantially different from the operations of the traditional banking (Cheng et al., 2006). There are copious empirical studies on the effect of web security and privacy on Internet banking acceptance. Perceived web security was found to favourably affect customers' intention to purchase on the Internet. The finding suggested that customers tend to purchase on the websites only if they believe that their credit cards and other sensitive information are safe (Centeno, 2003). Web security directly influenced the intention of banks' customers in Hong Kong to conduct banking transactions via the Internet (Cheng et al., 2006). Consumer security and privacy concerns were reported as an important potential barrier to Internet banking in some parts of Europe (Nasir, 2011). Further, perception of security of Internet banking also had a significant effect on customer intention to use Internet banking in Tunisia (Taylor and Todd, 1995). Conversely, website security and privacy did not significantly affect Internet banking acceptance in Finland (Pikkarainen et al., 2004). In the light of the aforementioned, we hypothesise that:

H4: Website security and privacy have a significant effect on user acceptance of Internet banking.

3.0 METHODS

Data for the study was collected via a research technique of survey conducted mid-September, 2019 at the Isolo campus of Lagos State Polytechnic. The survey was administered to the part-time students of Banking and Finance, Business Administration and Marketing. We collected and analysed properly filled 288 copies of questionnaire. There are five variables in the scale, four independent and one dependent, Internet banking acceptance. Two of the constructs of the independent variables, PU and PEOU were adapted from TAM (Davis, 1989). TAM is a widely used scale to predict or explain IS usage. In comparing TPB to TAM, Taylor and Todd (1995), noted that TAM is a better theoretical model for predicting usage. The last two constructs of this study, PE and web security & privacy, were adapted from (Pikkarainen et al., 2004). Each of the five constructs had four items, making a total of 20 items of measurement.

The measurement was done along a 5-point Likert scale, 1-5. 1 = Strongly disagree; 2 = Disagree; 3 = Undecided; 4 = Agree; and 5 = Strongly agree. A pilot study was earlier conducted on a sample size of 33 respondents. This enabled us to determine the Cronbach’s alpha for the scale. All Cronbach's alpha values exceeded the commonly adopted threshold value of 0.70 (Cronbach, 1951), suggesting the internal consistency of the items in the scale. This is shown in table 2 below. Cronbach's alpha coefficients did not exceed 0.95, suggesting that items are not redundant. Finally, the four hypotheses of the paper were tested with a multiple regression analysis.

4.0 RESULTS AND DISCUSSION

4.1 Descriptive Statistics

Table 4.1 Demographics of the respondents

Variables	Freq-uency	Percen-tage	Variables	Freq-uency	Percen-tage
Gender:			**Department**:		
Male	128	44	Banking & Finance	118	41
Female	160	56	Business Admin.	96	33
Total	288	100	Marketing	74	26
			Total	**288**	**100**
Age:			**Bank:**		
Below 20 yrs	60	21	GT Bank	87	30
21-30 years	185	64	Access/Diamond	46	16
31-40 years	42	15	UBA	26	9
41-50 years	1	0.3	Zenith Bank	24	8
Total	288	100	First Bank	21	7
			Others	88	30
			Total	288	100
Marital status:					
Single	224	76			
Married	64	22			
Separated	4	2			
Total	288	100			

Source: Source: Fieldwork, 2019.

The descriptive statistics shows that 56% of the respondents were female while the rest, 44% were female. The majority of the respondents, 64%, fell into the age category of 21-30 years. A large percentage of the respondents, 76%, were single. The frequency distribution of the respondents' department shows that 41% were students of Department of Banking and Finance, 33%, Department of Business Administration and 26%, Department of Marketing. Finally, the respondents' bank distribution indicates that 30% were customers of GT Bank; 16%, customers of Acces/ Diamond Bank; 9%, customers of United Bank of Africa; 8%, customers of Zenith Bank; 7%, customers of First Bank and the rest, 30% were customers of other banks in the country.

4.2 Reliability of the Sscale

Table 2: Reliability statistics

Constructs	No. of items	Cronbach's alpha
Internet banking usage	4	.905
Perceived usefulness (PU)	4	.888
Perceived ease of use (PEOU)	4	.873
Perceived enjoyment (PE)	4	.880
Web Security and privacy	4	.912
Entire scale	20	.911

Source: Fieldwork, 2019.

Table 2 shows the reliability for each of the five constructs of the paper and the aggregate scale for the entire scale. While the individual scale ranged from .873 to .912, the scale for the entire scale was .911. The results suggest that the reliability of the scale is above the .70 threshold specified by Cronbach (1952), suggesting that the scale is a reliable measure.

4.3 Hypotheses testing

Table 3: Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.731^a	.534	.528	2.613	1.651

a. Predictors: (Constant), Security and privacy of internet banking, Perceived usefulness of internet banking, Perceived enjoyment of internet banking, Perceived ease of use of internet banking

b. Dependent Variable: Internet banking usage.

Source: SPSS Output, 2019.

The R in table 3 is the *multiple correlation coefficients*. It determines the quality of the prediction of the dependent variable, Internet banking acceptance in this case. The value of R is .731 and suggests a good level of prediction. The R^2 value in the table is called the *coefficient of determination* and describes the amount of variance in the dependent variable that can be explained by all the independent variables combined together. Therefore, the (R^2 = 0.534) indicates that 53.4 per cent of the variance in Internet banking acceptance (dependent variable) was explained by the combination of the four independent variables of this paper – perceived usefulness of Internet banking, perceived ease of use of Internet banking, perceived enjoyment of Internet banking and web security and privacy.

Table 4: ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2218.888	4	554.722	81.214	.000^b
	Residual	1932.987	283	6.830		
	Total	4151.875	287			

a. Dependent Variable: Internet banking usage

b. Predictors: (Constant), Security and privacy of internet banking, Perceived usefulness of internet banking, Perceived enjoyment of internet banking, Perceived ease of use of internet banking.

Source: SPSS output, 2019.

The F-ratio in table 4 tests the extent to which the overall regression model is a good fit of the data. The results shows that at least one of the four independent variables – perceived usefulness of Internet banking, perceived ease of use of Internet banking, perceived enjoyment of Internet banking and web security and privacy - has a statistically significant effect on the dependent variable, Internet banking acceptance, i.e. F(4, 283) = 81.214, p < .0005, indicating that the model is a good fit. This suggests a linear relationship among the variables and that there was a 95% chance that this relationship is not due to chance.

In testing the four hypotheses stated in the paper, we used the value of unstandardised B and t-test statistics level of significance (Sig.) for each of the four factors examined in the coefficients table above. As a rule, where p<.05, we accept the hypothesis because it is statistically significant, otherwise we reject it. We report the results of the paper below.

Table 5: Coefficients[a]

Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	3.857	.661		5.839	.000	2.556	5.157
PU of Internet banking	.445	.066	.491	6.703	.000	.315	.576
PEOU of Internet banking	.160	.083	.168	1.917	.056	-.004	.324
PE of Internet banking	.056	.071	.059	.790	.430	-.083	.195
Security of Internet banking	.066	.052	.072	1.271	.205	-.036	.167

a. Dependent Variable: Internet banking usage

Source: SPSS output, 2019.

H1: Perceived usefulness of Internet banking has a significant effect on Internet banking acceptance. The results of hypothesis 1 indicates that perceived usefulness was statistically significant and had a positive effect on Internet banking acceptance for our sample (β = ·445, p<·05). Therefore, hypothesis 1 is supported, suggesting that perceived usefulness of Internet banking had a significant and positive effect on Internet banking acceptance.

H2: Perceived ease of use of Internet banking has a significant effect on Internet banking acceptance.

The results of hypothesis 2 shows that perceived ease of use of Internet banking did not statistically significantly influence Internet shopping for our sample (β = ·160, p>·05). Therefore, hypothesis 2 was not supported, suggesting that perceived ease of use of Internet banking did not have a significant effect on Internet banking acceptance.

H3: Perceived enjoyment of Internet banking has a significant effect on Internet banking acceptance.

The results of hypothesis 3 indicates that perceived enjoyment was statistically significant and had a positive effect on Internet shopping for our sample (β = ·056, p<·05). Therefore, hypothesis 3 was confirmed, suggesting that perceived enjoyment of Internet banking had a significant effect on Internet banking acceptance.

H4: Web security and privacy have a significant effect on Internet banking acceptance.

The results of hypothesis 4 indicates that web security and privacy were statistically significant and had a positive effect on Internet shopping for our sample (β = ·066, p<·05). Therefore, hypothesis 4 was also confirmed, suggesting that web security and privacy had a significant effect on Internet banking acceptance.

4.4 Discussion

We ran a multiple regression analysis to examine the factors that affect Internet banking acceptance in the context of Nigeria. Three of the hypotheses of the paper were supported while one was not supported. PU was found to significantly affect Internet banking acceptance. PU emerged as the strongest of the three factors that positively influenced Internet banking acceptance with a β value of ·445. This means that perceived usefulness makes the highest contribution, 44.5%, towards Internet banking acceptance in Nigeria. This result corroborates the results of past studies that relate perceived usefulness to user attitude, behavioural intention and actual usage (Pikkarainen et al., 2004; Cheng et al., 2006; Jahangir and Begum, 2008; Al-Smadi, 2012; Nath et al., 2014). This finding is not surprising because perceived usefulness of a new technology is an antecedent of its usage by consumers. Gardner and Amoroso posited that PU is considered to be more important than PEOU because users of a new technology would first consider its usefulness before getting to know how to use it and accepting its usage (Gardner and Amoroso, 2004). Similarly, our findings indicate that PE also significantly affects Internet banking acceptance. Though the contribution of PE to variance in Internet banking acceptance is lower, with a β value of .056, i.e. 5.6%, it is an important factor that drives Internet banking acceptance in Nigeria. This finding is consistent with the findings of previous studies (Teo et al., 1998; Bashir and Madhavaiah, 2015; Hormbaek and Hertzum, 2017). Our finding is at variance with the findings of (Pikkarainen et al., 2004; Suki and Suki, 2011), who found that PE did not significantly affect Internet banking acceptance. The supported finding suggests that the users of Internet banking have positive and reinforcing experiences most time they are transacting online. This category of users sees shopping, whether in the physical stores or online stores as a leisure activity through which they must maximise their time and experience. Suki and Suki (2012) explained that a user of a new technology will repeat an enjoyable activity while shunning a boring activity.

Moreover, the result on the effect of web security and privacy shows that this construct significantly affected Internet banking acceptance. Security and privacy construct contributed, with a β value of .066, 6.6% towards Internet banking acceptance. Again, this study confirmed the finding of past research (Salisbury et al., 2001; Centeno, 2003; Cheng et al., 2006; Nasir, 2011). Conversely, the result is contrary to the finding of (Pikkarainen et al., 2004). Usually, security threat and privacy risk constituted greater barriers to online shopping (Booz, Allen & Hamilton, 1997).

Consequently many consumers would avoid purchasing online if they believe that the online environment is not secure and cannot guarantee privacy of their personal data. Instructively, our finding suggests that the users of Internet banking in our dataset are secure and free of privacy risk and therefore, have no hesitation to conduct their banking transactions online. Cheng et al. (2006) affirmed this when they argued that customers expect to be secured and comfortable with Internet banking because of the unique ways its operations are carried out.

Finally, our empirical analysis revealed that PEOU does not have a significant effect on Internet banking acceptance. The finding is inconsistent with the findings of (Jahangir and Begum, 2008; Al-Smadi, 2012; Yoon and Occeña, 2014). However, the finding supported the findings of (Pikkarainen et al., 2004; Cheng et al., 2006). This finding is surprising because our*a prori* expectation was that Internet banking acceptance was positively affected by perceived ease of use of Internet banking. This finding means that Internet banking acceptance and usage by our respondents was not due to the ease of use of Internet banking. This result suggests that our sample would have still adopted the Internet for their banking transactions even if they Internet technology as complex to use. The users were the were more concerned usefulness, enjoyment and the security of the new and growing channel of banking services delivery. Importantly, the users considered usefulness as the most important facilitator of Internet banking transactions. Gardner and Amoroso affirmed that PU was more important than PEOU because users of a new technology would first consider its usefulness before finding out how and accept to use.

This paper has provided fresh evidence on the user acceptance of Internet banking in Nigeria, thus contributing to extant knowledge in this emerging research area in Nigeria and developing countries, at large.

5.0 CONCLUSION AND IMPLICATIONS

With the explosive growth in the population of Internet users in Nigeria, an increasing number of Nigerians are adopting the Internet for commercial purposes. Like in other countries, the Internet is gaining popularity as an alternative channel for delivering financial services in Nigeria. We extended the original TAM’s constructs - perceived usefulness and perceived ease of use of a new technology by two additional constructs – perceived enjoyment and security & privacy of the technology. We ran a multiple regression analysis to examine the effect of these four factors on Internet banking acceptance using student data collected from one of the foremost polytechnics in Nigeria, Lagos State Polytechnic. Three of the hypotheses formulated in the paper were supported by the data while the last hypothesis was not supported and these results are consistent with the earlier studies on Internet banking acceptance in other countries. Specifically, perceived usefulness, perceived enjoyment and security & privacy of Internet banking were found to significantly affect Internet banking acceptance. Surprisingly, perceived ease of use was found not to have a significant effect on Internet banking in Nigeria. Based on these findings, we therefore conclude that perceived usefulness, perceived enjoyment and security & privacy of Internet banking are significantly related to Internet banking acceptance and usage in Nigeria. Unlike perceived ease of use of Internet banking, our results affirm the importance of these three factors in motivating Internet banking acceptance in the context of Nigeria.

It is important to discuss the implications of our findings for Internet banking marketers in Nigeria. First, the result on perceived usefulness of Internet banking implies that the most important driver of Internet banking usage is the ability of Internet banking to facilitate smooth conduct of online banking transactions. Therefore, e-banking marketers should continue to enhance the functionality of their websites and emphasise this enhanced capability in their advert and other marketing communications to increase the adoption rate of Internet banking. Second, our result on perceived enjoyment of Internet banking suggests that Internet banking users expect some fun and enjoyment as they interact with the banks' websites for commercial purposes. The e-marketers should, therefore, enhance the entertainment capability of their websites to make user online experience less boring and more pleasant and enjoyable and motivate more fun-prone, youthful segment to accept Internet banking. It is believed that a user of a new technology will repeat an enjoyable activity while shunning a boring activity (Suki and Suki, 2011). Third, the finding on web security and privacy also suggests that these factors are paramount to Internet banking users in Nigeria. This finding implies that e-marketers should continue to fortify their website security to prevent criminals from gaining access to and stealing customers’ data. To allay the fears of potential Internet bank users and build trust and confidence in them, the e-marketers should widely promote the use of their more credible mobile applications, which are more secure than open Internet transactions using their URL via http://. Centeno (2003) posited that lower levels of trust in banks affect customer acceptance of Internet banking. Finally, Banks should frequently review and update and widely promote its privacy policies to assure and reassure existing and potential Internet banking users about the safety of the private information they share on the banks’ websites.

The following factors limit the extent to which the findings of this study are generalisable. First, a sample size of less than 300 is small for a study of this nature with millions of Internet banking users. Potential researchers in this area should consider a larger sample size. The study was carried out in Lagos State Polytechnic and the outcomes may not be generalisable to other students in Lagos state and Nigeria, at large. Future researchers should consider covering more campuses of tertiary institutions in Nigeria. Finally, the study excluded non-Internet banking users, whose population is larger than Internet banking users and this also limited the generalisability of the findings. Further research is required to validate the relationship between the factors examined in this paper and the intention of non-adopters to accept Internet banking.

REFERENCES

Agwu, E. M. and Carter, A. (2014). "Mobile Phone Banking in Nigeria: Benefits, Problems and Prospects." International Journal of Business and Commerce, 3(6), pp. 50-70.

Ajuwon, F. (2015). "Trends Set to Impact Nigeria's Banking Industry." Proshare. Available at: http://www.proshareng.com [Accessed 12 July 2019].

Ajzen, I. (1985). "From Intentions to Actions: A Theory of Planned Behaviour." Action control, pp. 11-39.

Al-Smadi, M. O. (2012). "Factors Affecting Adoption of Electronic Banking: An Analysis of the Perspectives of Banks' Customers." International

Journal of Business and Social Science, 3(17), pp. 294-308.

Aminu, S. A. (2012). "Empirical Investigation of the Effect of Relationship Marketing on Bank’s Customer Loyalty in Nigeria." Interdisciplinary Journal of Contemporary Research in Business, 4(6), pp. 1249-1266.

Aminu, S. A. (2013). "Challenges Militating against Adoption of Online Shopping in Retail Industry in Nigeria." Journal of Marketing Management, 1(1), pp. 23-33.

Aminu, S. A. (2018). "Technology Acceptance Model and Motorists' Intention to Adopt Point of Sale Terminals for Payment of Petrol Price in Lagos State, Nigeria." Governance and Management Review, 3(1), pp. 14-27.

Aminu, S. A., Oyefesobi, O. O. and Akerejola, W. O. (2018). "Survey of User Acceptance of Mobile Shopping in Nigeria." Proceedings of AICTTRA 2018 Conference, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria, pp. 132-139.

Aminu, S. A., Olawore, O. P. and Odesanya, A. E. (2018). "Perceived Financial and Privacy Risk Impediments to Internet Shopping." Proceedings of AICTTRA 2018 Conference, from November 11 to 14 at Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria, pp. 161-168.

Bashir, I and Madhavaiah, C. (2015). "Consumer Attitude and Behavioural Intention towards Internet Banking Adoption in India." Journal of Indian Business Research, 7(1), pp. 67-102.

Booz, Allen & Hamilton (1997). "Internet Banking: A Global Study of Potential." New York: Booz, Allen & Hamilton Inc.

Centeno, C. (2003). "Adoption of Internet Services in the Enlarged European Union: Lessons from the Internet Banking Case." Report EUR 20822, pp. 1-44. Available at: http://www.jrc.es [Accessed 30 June 2019].

Cheng, T. C. E., Lam, D. Y. C. and Yeung, A. C. L. (2006). "Adoption of internet banking: An Empirical Study in Hong Kong." Decision Support Systems, 42, pp. 1558-1572.

Chiemeke, S. C., Evwiekpaefe, A. E. and Chete, F. O. (2006). "The Adoption of Internet Banking in Nigeria: An Empirical Investigation." Journal of Internet Banking and Commerce, 11(3), pp. 1-9.

Cronbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests." Psychometrika, 16, pp. 297-334.

Davis, F. D. (1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." MIS Quarterly, 13(3), pp. 319-334.

Davis, F. D., Bagozzi, R. P. and P. R. Warshaw (1989). "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models." Management Science, 35, pp. 982-1003.

Davis, F. D., Bagozzi, R. P. and Warshaw, P. R. (1992). "Extrinsic and Intrinsic Motivation to Use Computers in the Workplace." Journal of Applied Social Psychology, 22, pp. 1111-1132.

Egwali, A. O. (2008). "Customer Perception of Security Indicators in Online Banking Sites in Nigeria." Journal of Internet Banking and Commerce, 13(3), pp. 24-26.

Ezeoha, A. E. (2005). "Regulating Internet banking in Nigeria: Problems and challenges, Part 1." Journal of Internet Banking and Commerce, 10(3), pp. 1-5.

Featherman, M. S., Miyazaki, A. D. and Sprott, D. E. (2010). "Reducing Online Privacy Risk to Facilitate E-service Adoption: The Influence of Ease of Use and Corporate Credibility." Journal of Services Marketing, 24(3), pp. 219-229.

Fishbein, M. A. and Ajzen, I. (1975). "Belief, Intention and Behaviour: An Introduction to Theory and Research." Massachusetts: Addison-Wesley, Reading.

Forsythe, S. M. and Shi, B. (2003). "Consumer Patronage and Risk Perceptions in Internet Shopping." Journal of Business Research, 56, pp. 867-875.

Gardner, C. and Amoroso, D. L. (2004). "Development of an Instrument to Measure the Acceptance of Internet Banking by Consumers." Proceedings of 37th Annual Hawaii International Conference on System Sciences, pp. 1-10.

Gerrard, P., Cunningham, J. B. and Devlin, J. M. (2006). "Why Consumers are not Using Internet Banking: A Qualitative Study." Journal of Services Marketing, 20(3), pp. 160-168.

Hernández-Murillo, R., Llobet G. and Fuentes, R. (2010). "Strategic Online Banking Adoption." Federal Reserve Bank of St. Louis Working Paper 2006-058E, pp. 1-33. Available at: http://research. stlouisfed.org/wp [Accessed 30 June 2019].

Hormbaek, K. and Hertzum, M. (2017). "Technology Acceptance and User Experience: A Review of the Experiential Component in HCI." ACM Trans. Comput- Hum. Interact, 24(5), pp. 33-63.

Internet World Statistics (2019). "Internet Usage and Population Statistics: Africa 2019." Available at: internetworldstatistics.com [Accessed 16 September 2019].

Jahangir, N and Begum, N. (2008). "The Role of Perceived Usefulness, Perceived Ease of Use, Security and Privacy, and Customer Attitude to Engender Customer Adaptation in the Context of Electronic Banking." African Journal of Business Management, 2(1), pp. 32-40.

KPMG (2017). "Nigerian banks and Internet banking penetration." Available at: http://home.kpmg [Accessed 3 July 2019].

Lallmahamood, M. (1970). "An Examination of Individual's Perceived Security and Privacy of the Internet in Malaysia and the Influence of this on their Intention to Use E-commerce: Using an Extension of the Technology Acceptance Model." PhD Dissertation. Malaysian Institute of Management. Available at: www.icommerce-central.com [Accessed 21 September 2019].

Naiyi, Y. (2004). "Dimensions of Consumer's Perceived Risk in Online Shopping." Journal of Electronic Science and Technology of China, 2(3), pp. 177-182.

Nasir, W. (2011). "Factors Influencing the Adoption of Internet Banking in Tunisia." International Journal of Business and Management, 6(8), pp. 143-160.

Nath, R., Bhal, K. T. and Kapoor, G. T. (2014). "Factors Influencing IT Adoption by Bank Employees: An Extended TAM Approach." The Journal for Decision Makers, 38(4), pp. 83-96.

Oke, S. (2019). "Internet Banking in Nigeria: How it all Works Explained." Available at: www.oasdom.com [Accessed 14 August 2019].

Ongkasuwan M. and Tantichattanon, W. (2003). "A Comparative Study of Internet Banking in Thailand." Paper presented at the First National Conference on Electronic Business, Bangkok, pp. 24-25 October, 2002. Available at: http://www.ecommerce.or.th/ [Accessed 16 August 2019].

Oni, A. A. and Ayo, C. K. (2010). "An Empirical Investigation of the Level of Users' Acceptance of E-banking in Nigeria." Journal of Internet Banking and Commerce, 15, pp. 1-13.

Pikkarainen, T., Pikkarainen, K., Karjaluoto, H. and Pahnila, S. (2004). "Consumer Acceptance of Online Banking: An Extension of the Technology Acceptance Model." Internet Research, 14(3), pp. 224–235.

Qureshi, T. M., Zafar, M. K. and Khan, M. B. (2008). "Customer Acceptance of Online Banking in Developing Economies." Journal of Internet Banking and Commerce, 13(1), pp. 1-12.

Reid, R. and Brown, S. (1996). "I Hate Shopping! An Introspective Perspective." International Journal of Retail and Distribution Management, 24(4), pp. 4-16.

Rogers, E. M. (1995). "Diffusion of innovations," 4th ed. New York: The Free Press.

Salawu, R. O. and Salawu, M. K. (2007). "The Emergence of Internet Banking in Nigeria: An Appraisal." Information Technology Journal, 6(4), pp. 490-496.

Salisbury, W. D., Pearson, R. A., Pearson A. W. and Miller, D. W. (2001). "Perceived Security and World Wide Web Purchase Intention." Industrial Management & Data Systems, 101, pp. 165-176.

Sathye, M. (1999). "Adoption of Internet Banking by Australian Consumers: An Empirical Investigation." International Journal of Bank Marketing, 17(7), pp. 324-334.

Suki, N. M. and Suki, N. M. (2011). "Exploring the Relationship between Perceived Usefulness, Perceived Ease of Use, Perceived Enjoyment, Attitude and Subscribers' Intentions Towards Using 3G Mobile Services." Journal of Information Technology Management, XXII(1), pp. 1-7.

Tan M. and Thompson, D. H. (2000). "Factors Affecting Adoption of Internet Banking." Journal of the Association for Information Systems, 1(5), pp. 1-42.

Taylor, S. and Todd, P. A. (1995). "Understanding Information Technology Usage: A Test of Competing Models." Information Systems Research, 6(2), pp. 144-176.

Teo, T. S, H., Tan, M. and Wong, K. B. (1998). "An Empirical Study of Adoptors and Non-adopters of Internet in Singapore." Information & Management, 34(6), pp. 339-345.

Venkatesh, V., Morris, M. G., Davis, G. B. and Davis, F. D. (2003). "User Acceptance of Information Technology: Toward a Unified View." MIS Quarterly, pp. 425-478.

Yoon, H. S. and Occeña, L. (2014). "Impacts of Customers' Perceptions on Internet Banking Use with a Smartphone." Journal of Computer Information Systems 54(3), pp. 1-9.

A TWO FACTOR AUTHENTICATION PROTECTIVE SYSTEM FOR MANAGING USER LOGIN CREDENTIALS

Ekundayo H. A., *Aminu E. F. and Alabelewe O. R.
Department of Computer Science, Federal University of Technology, Minna, Niger State, Nigeria

*Email of Corresponding Author: *enesifa@futminna.edu.ng*

ABSTRACT
Password managers are critical software programs that users rely on to store precious and sensitive data from passport and safety numbers to banking passwords in a protective and secure manner. However, it is always practically difficult for user to save or memorize numerous numbers of login credentials available for various web services. Password managers address both protection and usability issues by eliminating users ' need to study and memorize big amounts of data for various numbers of services and applications of web. To achieve this proposed system in this paper, we implore a two-factor authentication along with advanced encryption scheme as methodology. Therefore, this paper aims to design a protective system by considering biometric and encrypted strong passcode to manage various user login credentials. In the end, the proposed system incorporates the storage on a server of password encryptions. Similarly, security and trust are attained through both customer-side encryption and decryption. Also, in this paper, we tackle the problem of countering dictionary attacks by further enhancing the schema. Therefore, the proposed system has the capacity to eliminate users' stress and difficulties of recalling login credentials of numerous web services at any point of use.

Keywords: Password manger, encryption, decryption, two-factor authentication, dictionary attack.

1.0 INTRODUCTION

Based on research owing to the geometric growing of data at every given time on web, internet-based password authentication is evidently no longer secure. Internet, coined from the word International Network, is a set of computer networks that communicate using the Internet Protocol (IP). According to Norris, *et al.* (2001), Internet is a global network that uses the Internet protocol to interact. When the number of services offered on the internet has continued to rise tremendously, the number of passwords a typical user is required to remember correspondingly increases, to the point where it is no longer possible for most people to remember a new, complex password, for each and every account. Typical users fix the password difficulty by either jotting down the password in their note somewhere or reusing the same password across all accounts. Unfortunately, both methods can trigger vulnerabilities in the password.

The concept of using passwords for user authentication on the internet is cost-effective for services and easily comprehensible for users. However, the key challenge for user is being able to choose a strong password for each service and never reuse it for another service. This is important for the security, but the wide usage of passwords on the internet makes it impossible for users to memorize the required amount of different strong passwords (Liou & Bhashyam, 2010). To diminish this conflict of having user-friendly and secure passwords on possible solution of storing some data locally on the user's device, which can be the passwords themselves or information to compute them.

Being the first effective form of computer-based authentication, passwords are increasingly becoming a security problem in the modern age. There are an increasingly number of websites emerging on the internet, each demanding its own username and password. A recent study reveals that internet users, on average, have about 25 accounts that require password protection (Conklin *et al.*, 2004).

The behavioural liability of choosing protective, alphanumeric passwords across all locations that rely heavily on password authentication is a major problem with password verification. A large body of reliable research gabs suggests users have potentially, sensibly given up, choosing simple passwords and reusing them across sites (Harley., 2009). Password managers are aimed at providing a way out of this unrelated scenario. A safe password manager could create and store passwords for customers automatically, distancing oneself from the cognitive burden of having to remember them.

A password Manager is one of the best ways to keep track of each unique password or passphrase that you have created for your various online accounts (e.g. Facebook account, Twitter account, Instagram account, Research Gate account, and other important account login credentials) without writing them down on a piece of paper and risking that others will see them (Li, Zhiwei *et al.,2014*) .At its core, A password manager operates as a safe storage to record user passwords and account usernames such as Facebook account, Twitter account, Instagram account details, bank account details and other significant account login credentials. The password manager uses a master username and password to secure access to this database. With a robust master password, a safe password manager guarantees that a user can depend on separate, unguessable passwords for each web service without the related cognitive burden of memorizing them all.

Instead, only one strong master password should be remembered by the user.

This paper aims to eliminate the bottlenecks associated with memorizing numerous login credentials accounts of various web services. Such as Email, Twitter, Instagram. A password manager does have a warehouse or memory on different web apps of the login details of a user. A web application is a website that verifies its users by demanding a mixture of username and password. The "entry point" of the web app is the section where the user of the application can enter his username and password. We call an entry point, username, and password mixture of a credential. For the same web application, a user can store various credentials, in which case name separates each credential.

1.1 Two Factor Authentications

Two-factor Authentication is a system in which authenticated combination of two different factors are used. Using two factors as compared to one factor usually results in a higher level of authentication assurance. In August 2006, the FFIEC released additional guidelines on this topic clarifying, "By default, true multifactor authentication involves the use of two solutions or more of the three categories of factors, using multiple solutions from the same category would not constitute multifactor authentication.”

This second factor typically takes the form of a user's physical security token or smart card. This is referred to as the factor of what you have. In this case, mobile phone and other personal devices may also be used by some application. One example is using any bank's issued ATM card. One authentication factor is the customer's physical ATM card that customer slides into the ATM machine. The second factor is the PIN the customer enters. Without both, authentication cannot take place.

Another application of the second factor, such as a fingerprint scan, may be a biological factor. This is referred to as the factor of what you are. Using the what-you-are factor requires special equipment to scan input data, which means that delivery costs and complexity are higher.

The research work of Liou and Bhashyam, (2017) improve on security, stated that the information in the what-you-have factor should be changed along the time. Thus, the information is no longer valid when it is stolen and re-used. This is called One Time Password.

2.0 LITERATURE REVIEW

There is significant literature work that has been done on how user information can be kept safe. Various researchers have contributed in this area but not without room for improvement. In view of this, Sandvoll *et al.,* (2014) designed and analyzed password management system. The password management system has been designed and implemented as an iOS application called PassCue. PassCue is based on the Shared Cues password management model, the design and implementation choices, as well as parameter evaluation, were important in order to create a usable and secure system. PassCue uses cues to share secrets across multiple accounts in order to achieve the competing usability and security goals.

In the work of Whitten *et al.*, (1999) that evaluated the usability of a security program, (Pretty Good Process) PGP 5.0, have concluded that usability principles should be extended beyond the ones commonly used for generic user interface evaluation and design. They claim that security as a field has a series of properties that introduce extra difficulties in the design of secure and usable systems.

Pretty Good Privacy (PGP) pretty good privacy is an encryption program that provides cryptographic privacy and authentication for data communication. PGP is used for signing, encrypting, and decrypting texts, e-mails, files, directories, and whole disk partitions and to increase the security of e-mail communications.

According to McCarney *et al*., (2012) design and implementation password manager application, named Tapas, that works with dual device authentication instead of a master password to encrypt the database. Their study aimed to alleviate the users from the burden of memorability completely while keeping a good usability standard as other popular password manager applications that they used to compare Tapas.

Many more studies for instance, Veras *et al*., (2014) have confirmed the frustration of users and its effects on password creation. They found that users would fulfill policy requirements in predictable ways such as use only a small fraction of the symbols on a keyboard, choose semantically meaningful passwords and password-phrases that follow grammatical rules.

Finally, another study on password-creation policies carried out by Shay *et al*., (2016) found that the usual comp8 is very susceptible to both online and offline attacks and should be replaced with more usable and secure alternatives, like the 2word16 or 2class12.

Table1 presented the summary of password managers related literature that consider storage resources available as well as the systems represented have been evaluated in specifics together with the database layout being used by software. In this paper, we argue if source code is obtainable and that the password manager is incorporated with a browser.

3.0 METHODOLOGY

3.1 The Proposed Protective System

The technique implore in this research work as methodology uses a two-factor authentication to address certain security issues while bringing in some feasibility issues and other security concerns. Therefore, in this section, we presented the protective system’ conceptual designs that describe how activities are being carried out by the user from the point where user registers account to the point where user information is being retrieved.

Table 1: Analysis of the existing and proposed system

S/No.	Related Work	Authors/Year	Databasse Layout	Storage Type	Open Source	Platform/Device	Browser Integration
1.	Design and analysis of Passcue password manager	Marts *et al.*, 2014	Microsoft Internet Explorer(MSIE)	Local/cloud	Yes	IOS	No
2.	Design and implementation of password manager (Tapass)	McCarney *et al.*, 2012	Knowledge of database	Local	Yes	Windows/Linux	No
3.	Development of a password manager (LastPass)	Karole *et al.*, 2011	Knowledge of database	Local/Cloud	No	Windows	Yes
4.	Design and analysis of a smart phone based password manager (KeePassMobile)	Shey *et al., 2010*	Knowledge of database/knowledge Of database X4	Local	Yes	Windows/Linux/Max	No
5.	The proposed System (dual possession authentication protective system for managing user login credentials)	Ekundayo *et al.*, 2019	Knowledeg of database	Local/Cloud	Yes	Windows	No

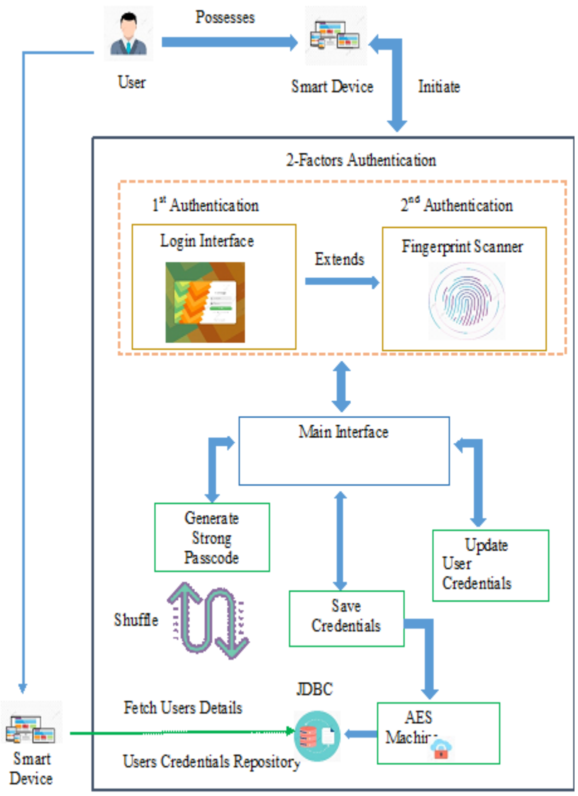


Figure 1 – System Architectural Overview of the Proposed System

From figure 1, a user uses their computer to initiates the application. It is expected of them to supply necessary information such as username and password, after which they will be authenticated using a finger-print recognition. After the authentication has been successful, user can now proceed in using the application. Users are allowed to perform some certain activities listed below.

(i) Initiates the system.
(ii) Pass through the authentication process without trespassing any security protocol.
(iii) Use the system to generate a very strong password.
(iv) Use the system to encrypt their personal credentials.
(v) Use the system to fetch or retrieve encrypted credentials that was stored.

3.2 Main Interface

The main interface will be presented to the user after the user have successfully registered their account. this interface provides user to be able to performs the following action.

Generate user with strong password

Generating strong password for user is having to do the work of combining user choice of words for their password, shuffle it together and their by using those input words to produce a very strong password that is very difficult to break or guess.

Save credentials

User information will be store locally in their system and cloud where user will be able to reach at any part of the word remotely

Update user credentials

Users information maybe out of date according to user's specification and it needed to be updated so as to meet the current trend at that time. One of the purposes that the main interface serves, is to allow user update their information as soon as possible. It is very important that user have already registered account.

AES Encryption engine

AES which is a short form of Advanced Encryption Standard is used in order to protect data against unauthorized access and to encrypt this. The cryptographic process key of varying length is utilized for this purpose. This is designated AES-128, AES-192 OR AES-256 depending on the length. The process was originally introduced by the American national institute of standards and technology and can be used in the USA to encrypt documents with a maximum security rating. This method of encryption of any type data is considered to be particularly secure and effective. This AES will encrypt user information that is supply into the system before saving it into the database so as to increase the cost of attacker.

JDBC Java Database Connectivity

JDBC stands for Java Database Connectivity, which is a standard Java Application Programming

Interface (API) for database independent connectivity between the java programming language and a wide range of databases.

The JDBC library includes APIs for each of the tasks mentioned below that are commonly associated with database usage.

(i) Making a connectivity to database.
(ii) Creating SQL or MySQL statements.
(iii) Executing SQL or MySQL queries in the database
(iv) Viewing & Modifying the resulting records.

Fundamentally, JDBC is a specification that provides a complete set of interfaces that allows for portable access to an underlying database. Java can be used to write different types executables, such as:

(i) Java Applications
(ii) Java Applets
(iii) Java Servlets
(iv) Java Server Pages (JSP)
(v) Enterprise JavaBeans (EJBs).

All of these different executables are able to use a JDBC driver to access a database, and take advantage of the stored data.

JDBC provides the same capabilities as Open Database Connectivity (ODBC) allowing java programs to contain database independent code

3.3 JDBC Architecture

The JDBC API supports both two-tier and three-tier processing models for database access but in general, JDBC Architecture consists of two layers-

(i) JDBC API: this provides the application TO-JDBC Manager connection

JDBC Driver API: This supports the JDBC Manager-to-Driver Connection.

The JDBC API uses a driver manager and database-specific drivers to provide transparent connectivity to heterogeneous databases.

The JDBC driver manager ensures that the correct driver is used to access each data source. The driver manager is capable of supporting multiple concurrent drivers connected to multiple heterogeneous database.

Figure 2 in addition implore Use Case Diagram to depict the scenarios functionality of the proposed System.

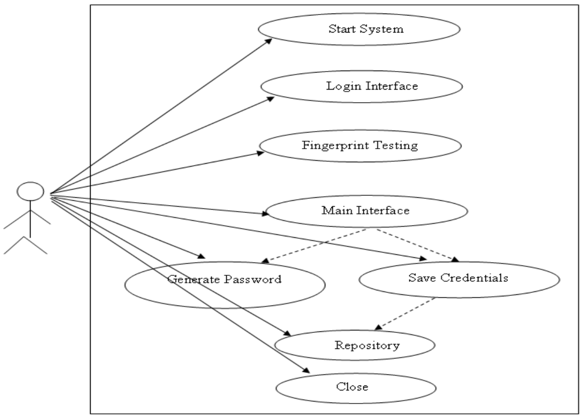


Figure 2: Use case of diagram of the proposed System

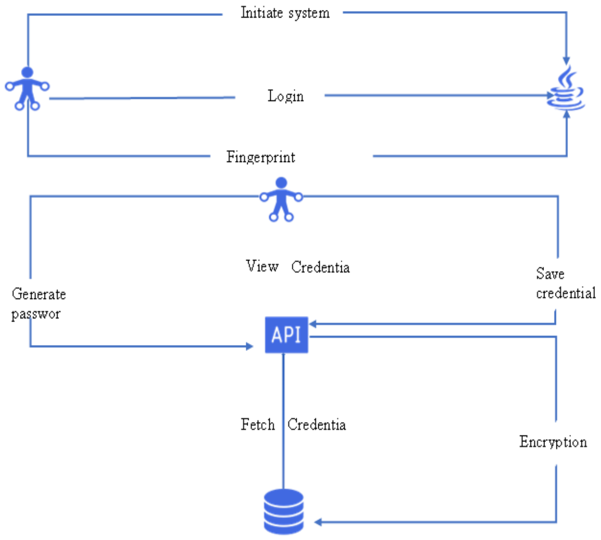


Figure 3: Data Flow Diagram The Proposed System

Figure 3 describes the data flow in the proposed system. Dataflow diagram is an easy to comprehend graphical method which aids in describing the boundaries of the system. It is beneficial for communicating the currents system data which describes the logic behind the data flow within the system to the user. It is also used to model a new system. The following are the advantages of the data flow diagram over other diagrams:

(i) Data flow diagram could visually "state" things that might be very difficult to describe in words, and they function for both technical and non-technical audiences.
(ii) They are much less relevant nowadays to visualizing interactive, real time or data base oriented software or system.
(iii) It could be used as an initial step to develop an overview of the system without entering deep detail.

4.0 RESULTS AND DISCUSSION

The data obtained by testing the password manager system was analysed to approve the proposed design. After a successful registration of user 's account. users are allowed to save their login details, generates a strong password for their multiple accounts. Also, share credentials through a secure channel.

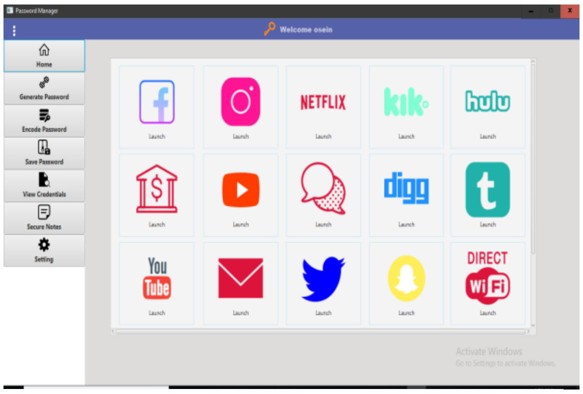
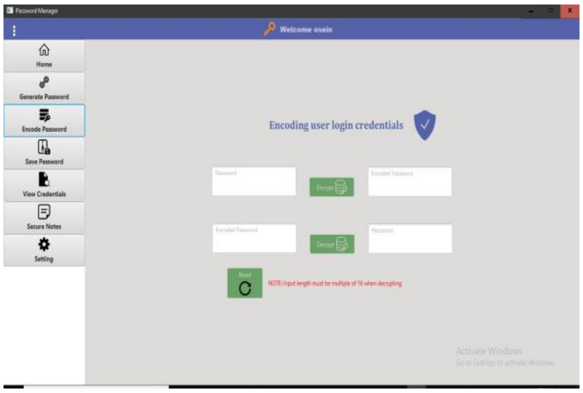


Figure 4: Encrypting user information

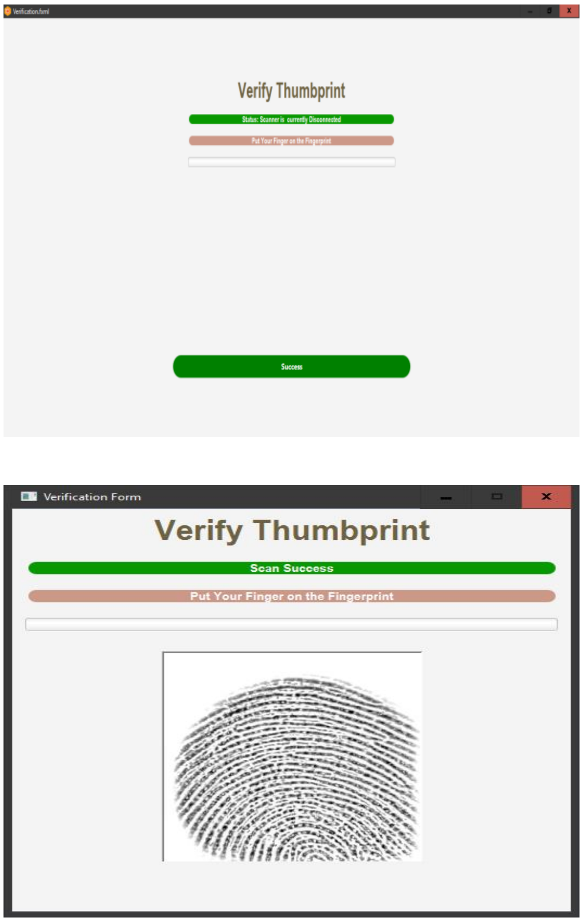


NOTE: Input length must be multiple of 16 when decrypting

Figure 5: Encrypting and decrypting mechanism

The main interface presented incorporates the activities perform by the proposed password manager called passhouse. By the left of the interface, there are six major components, Home, Generate Password, Encode Password, Save Password, View Credentials, Secure Notes.

When the user registers with the software service, it will obtain a username and passwords for the software service. When the user wishes to store the password for the password manager service on the storage, it must first encrypt them with the user encryption key before transmitting them. The encryption mechanism makes use of user own generated key for the encryption.



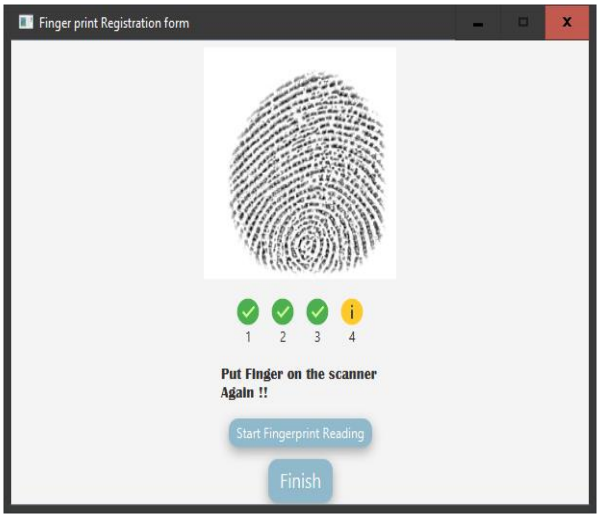
Status: Scanner is currently Disconnected

Put Your Finger on the Fingerprint

Scan Success

Put Your Finger on the Fingerprint

Figure 6: User verification using a fingerprint scanner



Put Finger on the scanner
Again !!

Figure 7: Fingerprint Scanning process

5.0 CONCLUSION

This project provides solution to the problem people are facing when it comes to remembering their login credentials details, and also of generating strong password for their digital account. The use of this system is limited to users or organizations with personal computer. The implemented system has the following features; fast navigation for user to perform tasks, well-built user interactive application, fast retrieval of user information, mechanism for generating a strong password for users. The proposed system is developed based on java technology and two factor authentications along with Advanced Encryptions Security algorithm is also enforced. Thus, in this paper, the proposed system possesses the features and strength to remove users' challenge of recalling login credentials of numerous web services at any point of use.

REFERENCES

Alkaldi, N., & Renaud, K. (2016). Why do People Adopt, or Reject, Smartphone Security Tools? *EuroUSEC 2016: The 1st European Workshop on Usable Security*, (July), 1–15. https://doi.org/10.14722/eurousec.2016.23011

Aurigemma, S., Mattson, T., & Leonard, L. (2017). *So Much Promise , So Little Use : What is Stopping Home End-Users from Using Password Manager Applications ?* 4061–4070.

Bharadwaj, A. S. (200). A resource-based perspective on information technology capability and firm performance: an empirical investigation. MIS quarterly, 169-196

Chaudhary, S., Schafeitel-Tähtinen, T., Helenius, M., & Berki, E. (2019). Usability, security and trust in password managers: A quest for user-centric properties and features. *Computer Science Review*, *33*, 69-90.

Conklin, Art, Glenn Dietrich, and Diane Walz. "Password-based authentication: a system perspective-based authentication: a system perspective." 37th Annual Hawaii international Conference on System Sciences, 2004. proceedings of the. IEEE, 2004

Gasti, P., & Rasmussen, K. B. (2018). *On The Security of Password Manager Database Formats*.

Hakbilen, O., Perinparajan, P., Eikeland, M., & Ulltveit-Moe, N. (2018). *SAFEPASS - Presenting a Convenient, Portable and Secure Password Manager*. (Icissp), 292–303. https://doi.org/10.5220/0006603102920303

Horsch, M., Andreas, H., & Buchmann, J. (2015). *PAsswordLess PAssword Synchronization Extended Version*.

Liou, J. C., & Bhashyam, S. (2010). A feasible and cost effective two-factor authentication for online transactions. *2nd International Conference on Software Engineering and Data Mining, SEDM 2010*, 47–51.

Maqbali, F. Al, & Mitchell, C. J. (2019). Web password recovery: A necessary evil? *Advances in Intelligent Systems and Computing*, *881*, 324–341. https://doi.org/10.1007/978-3-030-02683-7_23

Quelch, J. A., & Klein, L. R. (1996). The Internet and International Marketing. Sloan Management Review, 37(3), 60-70.

Sandvoll, M. B. (2014). *Design and analysis of a password management system* (Master's thesis, Institutt for elektronikk og telekommunikasjon).

Sandvoll, M., Boyd, C., & Larsen, B. B. (2014, December). PassCue: The Shared Cues System in Practice. In *International Conference on Passwords* (pp. 119-137). Springer, Cham.

Li, Z., He, W., Akhawe, D., & Song, D. (2014). The emperor’s new password manager: Security analysis of web-based password managers. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)* (pp. 465-479).

Sandvoll, M. B. (2014). *Design and analysis of a password management system* (Master's thesis, Institutt for elektronikk og telekommunikasjon).

Whitten, A., & Tygar, J. D. (1999, August). Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *USENIX Security Symposium* (Vol. 348, pp. 169-184).

FFIEC press release. Visited and retrieved on 28th September, 2019, http://www.ffiec.gov/press/pr081506.htm.

ONTOLOGY MODULARIZATION IN AFRICAN TRADITIONAL MEDICINE USING DATA CLUSTERING

[1]Awoniran O. M., [2]Soriyan H. A., [3]Elujoba A. A., [2]Ikono R. N. and [4]Okunnade A. A.
[1]Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria.
[2]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.
[3]Department of Pharmacognosy, Obafemi Awolowo University, Ile-Ife, Nigeria.
[4]Department of Physics and Engineering Physics, Obafemi Awolowo University, Ile-Ife, Nigeria.

**Corresponding author's e-mail: lekanawoniran@gmail.com*

ABSTRACT

This research work approaches ontology modularization from a multivariate data clustering approach. The methodology employed in the work firstly involves the extraction of the classes and attributes of the ontology. Thereafter, proximity measures were computed between the various concepts in the domain obtained from classes and attributes extraction. An agglomerative hierarchical clustering algorithm was then applied to the proximity measures and the resulting clusters of concepts were analyzed and validated. Results from the work show eight sets of clusters that were obtained at various levels of dissimilarity values (d = 0.0 to d = 0.7). By using the sum of squared errors for validation, the optimal cluster was found to be at the fourth dissimilarity value (d = 0.3). Conclusively, the multivariate data clustering approach provides an effective approach to ontology modularization. The results show that domain knowledge processes such as knowledge capture, knowledge representation, knowledge documentation and knowledge transfer of the concepts in the ATM practice should be done together as obtained in this research work in order to obtain holistic and wholesome results.

Keywords: Ontology, Ontology Modularization, African Traditional Medicine, Data Clustering.

1.0 INTRODUCTION

Ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that may really or fundamentally exist in a particular domain of discourse [1]. It is a description of the ideas, concepts and relationships that exist in a domain and it compartmentalizes the variables needed for some set of computations and establishes the relationships between them [2]. The result of an ontology creation process is usually a large database of semantic data. In large domains, scalability, computability, and applicability of ontologies could be quite difficult. The modularization of such ontologies enhances these factors. Ontology modularization structures ontologies into groups of independent and re-usable knowledge sub-sections [3]. An example of a large domain is African traditional medicine (ATM) practice [4].

The ATM practice consists of practices (both explicable and inexplicable), for the diagnosis, prevention and elimination of physical, mental and social disequilibrium. It is usually handed down verbally from generation to generation by a community of people in African society [5]. The various aspects of the ATM practice as well as the various medicinal plants used for the treatment or management of various diseases give rise to a large and complex domain. Some aspects of the practice include orthopaedics, psychiatry, midwifery and general medical practices [6]. On the part of the medicinal plants, various medicinal properties can be attributed to specific medicinal plants. Thus, relating illness to medicinal plants will give a many-to-many relationship. This is because most medicinal plants possess multiple medicinal properties while most illnesses require multiple medicinal properties for their effective treatment.

To resolve this ambiguity, the illnesses can be attributed to the medicinal properties of the plants by which they are treated. However, a multivariate set of data will be the result of so doing. Therefore, clustering this data will result in a basis for the modularization of ontology in the ATM domain.

This research aims at clustering the multivariate set of data that arises by attributing the medicinal properties of plants used in the treatment and/or management of some illnesses to the therapies of the illnesses in the ATM practice. It focusses on establishing relationships between various illnesses through the medicinal properties of the herbs that are used in the treatment and/or management of such illnesses.

This paper is structured as follows; Section 1.0 gives an introduction and background to the research; Section 2.0 reviews existing works that are related to multivariate data clustering and ontology modularization. Section 3.0 discusses the research methodology the methods employed in carrying out the research work. Section 4.0 shows the results obtained from the work and discusses the implications of the results. Finally, section 5.0 concludes the research work.

2.0 LITERATURE REVIEW

An algorithm for the clustering of multivariate data streams with correlated components was proposed for the clustering of datasets from large databases that arrive in streams [7]. The authors identified that existing clustering algorithms do not consider the specific cluster variability or assume that data have uncorrelated components and can be plotted as random vectors. The methodology employed in the work involves the creation of covariance matrices through the double shrinkage method. Afterward, distance matrices were then created using the Mahalanobis distances for the data clustering. The developed algorithm was applied to the detection of a network. Results from the work show a stable cluster of data with correlated components. In situations that might have resulted in instability, Mahalanobis distances provided the necessary error correction to achieve a good degree of stability and clusters with high precision [6]. However, the method data streaming. Hence, the application of the method to unstructured data will result in complexities.

A technique for improving semantic image annotation through the use of ontology modularization was proposed [8]. The aim of the technique is to solve the problem of insufficient details in medical image annotations. Naming of medical images is often not described in details thus making it difficult to utilize the inherent wealth of information contained within the images. A solution to this is the annotation of the data contained within the images with concepts from medical ontologies. However, this is an enormous task due to the size and complexity of medical ontologies [8]. In order to reduce the size of the clinical ontologies for a more efficient image and text annotation, the modularization of the ontologies was proposed. The methodology adopted in the work involves the identification of smaller fragments of a large anatomy ontology that are relevant for annotating medical images from patients suffering from a named disease. For each new context, a new representative corpus was then created by which the annotations and modularizations are carried out. Results from the works show a grouping of diseases by which image annotations are carried out. This reduces the complexities encountered in the annotation of medical images. However, a major drawback of the work is that the approach taken produced a rather coarse-grained set of ontology modules which reduces the specificity of the medical image annotations. This will give rise to a broader set of image annotations.

Ontology partitioning by data clustering was proposed as a method for achieving ontology modularization [9]. The proposed technique aims at improving data sharing, matching, reuse, integration as well as the scalability. The necessity for the creation of a technique for achieving this arose due to the increased use of the semantic web. This has increased the representation of knowledge as semantic data and consequently resulted in an increased construction of ontologies. The methodology adopted in the methodology involves using an enhanced k-means clustering algorithm based on proximity measures (the similarity matrix) for partitioning an ontology into modules. Results from the work show that employing the Dennai's measure in the computation of the proximity measures resulted in high-quality clusters after the application of the clustering algorithm. However, a major drawback of the work is that it only considers ontologies that could be represented as univariate or bivariate data. It does not cater for complex ontologies that are represented as multivariate data. Also, clarity was not provided as to the effective means of representing ontologies on a distance graph.

A methodology for knowledge handling in ontology-based information extraction systems was also proposed [10]. It aims at reducing the complexities encountered in ontology construction, updating, extraction, positioning, and sharing. The method adopted in involves the application of ontology-based information extraction tools for the purpose of ontology construction. Results from the work show large ontologies constructed with fewer complexities using ontology-based extraction tools. The methodology provides a method of easy construction of particularly large ontologies. However, the specificity of the developed ontology is quite low.

Currently, methods for achieving ontology modularization through the clustering of data are majorly channeled towards univariate and bivariate data. In situations where multivariate data is involved, Mahalanobis distances are computed which converts the multivariate data to bivariate data.

3.0 METHODOLOGY

The methodology employed in the research work involves firstly, the extraction of the classes and attributes of the targeted concepts. Thereafter, proximity measures were computed and the concepts were then clustered. The methodological framework is shown in figure 3.1.

3.1 Classes and Attributes Extraction

The classes and attributes extraction is the identification and extraction of classes and attributes from the ontology. For the research, the classes and attributes pertaining to illnesses and their therapies were focused on. Data obtained from the research work on the establishment of consensus in the knowledge base creation of medicinal plants of African traditional medicine practice [11] were used for the classes and attributes extraction.

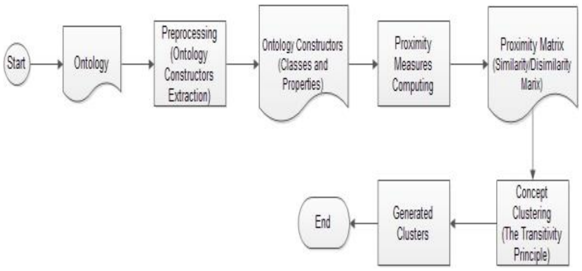

Start
Ontology
Preprocessing (Ontology Constructors Extraction)
Ontology Constructors (Classes and Properties)
Proximity Measures Computing
Proximity Matrix (Similarity/Disimilarity Marix)
Concept Clustering (The Transitivity Principle)
Generated Clusters
End

Figure 3.1: The Methodological Framework

3.2 Proximity Measures Computing

Proximity measure is the measure of similarity or dissimilarity between concepts [12]. For the research, the proximity measures of concepts were computed using the therapies employed in the treatment of various illnesses. The measures (similarity and dissimilarity) provides a means for the recognition of patterns in the concepts obtained in the previous sub-section.

In carrying this out, firstly, the various attributes relating to therapies provided for various illnesses were identified and put in an array. Using [1: n] array. The set of attributes (A) is given as

$$A = [a_0, a_1, a_2 \dots a_{n-1}] \qquad 1.1$$

Also, the extracted concepts (C) (which are also regarded as classes) are equally put into an array of [1: m] as shown below;

$$C = [C_0, C_1, C_2 \dots C_{m-1}] \qquad 1.2$$

Thereafter, a table of concepts (inputted as the rows) and attributes (inputted as columns) was computed. Each record in the table consists of a concept C with its set of attribute values. For a concept C_y such that C_y ε C, the class C_y is written as;

$$C_y = [a_0, a_1, a_2 \dots a_{n-1}] \qquad 1.3$$

However, null values were recorded for cases where an attribute cannot be related to a particular concept. Afterward, the dissimilarities between the concepts were computed. Given two concepts C_a and C_b, the dissimilarity between the concepts can be mathematically written as;

$$d(C_a, C_b) = 1 - \frac{\text{Number of matches}}{\text{Number of attributes}} \qquad 1.4$$

Dissimilarity measures were computed for all concepts in the ontology and represented as an *n* x *n* matrix. This is regarded as the dissimilarity matrix (which is the direct opposite of the similarity matrix).

3.3 Clustering of the Concepts

For the research, the clustering of concept was carried out such that given an associated set of dissimilarity measures, a cluster of homogenous concepts is a set of concepts with the minimum value of dissimilarities.

Mathematically, given a set of dissimilarities, $D_{ca,}$ for an entity C_a (a concept) as [d(C_a, C_b), d(C_a, C_c), d(C_a, C_d) … d(C_a, C_m)], Cluster K_a with maximum homogenous relation and minimum heterogeneous relation to entity C_a is*;*

$$\min[D_{ca}] \qquad 1.5$$

This can be further written as;

$$min[d(C_a, C_b), d(C_a, C_c), d(C_a, C_d) \dots d(C_a, C_m)] \; 1.6$$

Furthermore, a denser cluster was achieved by adding clusters with minimum dissimilarity values in which the concepts that were clustered belonged to originally Cluster K_a will then be given as;

$$min[D_{ca}] \wedge min\ [min[D_{ca}]] \qquad 1.7$$

This also can be further written as;

$$min[d(C_a, C_b), d(C_a, C_c), d(C_a, C_d) \dots d(C_a, C_m)]$$
$$min\ [min[d(C_a, C_b), d(C_a, C_c), d(C_a, C_d) \dots d(C_a, C_m)]] \qquad 1.8$$

By adding more layers of minimum dissimilarity values of entities to the cluster, denser clusters were achieved. The algorithm below shows the clustering process.

Algorithm: Concept Clustering Algorithm

```
INPUTS:
Dissimilarity_set, $D_{cm}$
Number_of_dissimilarity_sets, m
Minimum dissimilarity value, d
OUTPUTS:
Clusters
BEGIN:
n = 1
While (n ≤ m)
d = 0.1
$\min[D_c]$ = d
Add to cluster
d = 0.1+0.1
End While
End
```

3.4 Cluster Analysis and Validation

Various sets of clusters were obtained after the clustering process. Therefore, obtaining the optimal cluster of concepts through cluster analysis and validation is important. The aim of cluster analysis and validation is to find groups of objects (array of concept clusters) such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The within-cluster sum of squares (WSS) was used in achieving this. WSS was computed for the different various clusters using the formula below [13]*;*

$$\text{WSS} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad 3.12$$

$$\text{WSS} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2 \qquad 3.13$$

Where $\bar{x}$ is the mean of the dissimilarity which is measured in each cluster and x is the individual dissimilarity measure.

4.0 RESULTS AND DISCUSSION

The results obtained at the various stages of the research work are provided in the following subsections.

4.1 Classes and Attributes

"Illness" was considered as a class and "medicinal properties of plants" was considered as an attribute. Various instances of the "illness" class and their "medicinal properties" attributes are given in Table 4.1. Also, the various medicinal attributes of medicinal plants are coded as given in Table 4.2.

Table 4.1: Illnesses and Medicinal Properties Instances

S/No.	Class	Medical Term	Properties of Medicinal Herbs used for Treatment/Management of Ailments
1	C1	Asthma	Analgesic, Anti-allergic, Anti-asthmatic, Antimicrobial, Demulcent, Relaxant.
2	C2	Chicken Pox	Anti-allergic, Antimicrobial, Antipruritic, Antipyretic, Antiseptic, Emollient.
3	C3	Cold & Catarrhal	Analgesic, Anti-allergic, Antimicrobial, Demulcent, Relaxant.
4	C4	Convulsions	Adaptogenic, Analgesic, Anti-anaemic, Anti-inflammatory, Antispasmodic, Relaxant.
5	C5	Cough	Anti-allergic, Antimicrobial, Demulcent, Relaxant.
6	C6	Diabetes	Anti-hyperglycaemic, Anti-coagulant, Antioxidant, Nutritive/Tonic.
7	C7	Diarrheal	Alkaline, Anthelminthic, Antidiarrheal, Antimicrobial, Astringent, Nutritive/Tonic.
8	C8	Eczema	Anti-allergic, Antimicrobial, Antipruritic, Antiseptic, Emollient.
9	C9	Fibroid	Anti-hyperglycaemic, Anti-inflammatory, Anti-tumor, Diuretic, Emmenagogue, Nutritive/Tonic.
10	C10	Gonorrhea	Analgesic, Antimicrobial, Anti-inflammatory**,** Antipyretic, Anti-tumor, Aphrodisiac, Diuretic, Nutritive/Tonic.
11	C11	Hernia	Adaptogenic, Analgesic, Astringent, Relaxant.
12	C12	Hypertension	Anti-coagulant, Anti-hyperglycaemic, Antioxidant, Hypotensive, Nutritive/Tonic.
13	C13	Erectile Dysfunction	Antimicrobial, Aphrodisiac, Diuretic, Relaxant, Nutritive/Tonic.
14	C14	Malaria	Adaptogenic, Analgesic, Anti-anaemic, Anti-emetic, Anthelminthic, Antimalarial, Antimicrobial, Antipyretic, Nutritive/Tonic
15	C15	Pile	Alkaline, Antimicrobial, Antidiarrheal, Anthelminthic, Astringent, Nutritive/Tonic, Relaxant
16	C16	Rheumatism	Analgesic, Anti-anaemic, Anti-coagulant, Antipyretic, Nutritive/Tonic, Relaxant.
17	C17	Ring Worm	Antimicrobial, Anti-inflammatory, Antipruritic, Antiseptic, Emollient.
18	C18	Stomach Disorder	Alkaline, Anthelminthic, Antimicrobial, Astringent, Nutritive/Tonic.
19	C19	Syphilis	Analgesic, Antimicrobial, Anti-inflammatory**,** Antipyretic, Anti-tumor, Aphrodisiac, Astringent, Diuretic, Nutritive/Tonic.
20	C20	Typhoid	Adaptogenic, Analgesic, Anti-anaemic, Anti-emetic, Anthelminthic, Antimicrobial, Antipyretic, Astringent, Nutritive/Tonic.
21	C21	Ulcer	Alkaline, Anthelminthic, Antimicrobial, Astringent, Relaxant
22	C22	Yellow Fever	Adaptogenic, Analgesic, Anthelminthic, Anti-anaemic, Antimicrobial, Antipyretic, Diuretic, Nutritive/Tonic.

Table 4.2: Medicinal Attributes

S/N	Code	Attributes
1	a_1	Adaptogenic
2	a_2	Alkaline
3	a_3	Analgesic
4	a_4	Anthelminthic
5	a_5	Anti-allergic
6	a_6	Anti-anaemic
7	a_7	Anti-asthmatic
8	a_8	Anti-coagulant
9	a_9	Antidiarrheal
10	a_{10}	Anti-emetic
11	a_{11}	Anti-hyperglycaemic
12	a_{12}	Anti-inflammatory
13	a_{13}	Antimalarial
14	a_{14}	Antimicrobial
15	a_{15}	Antioxidant
16	a_{16}	Antipruritic
17	a_{17}	Antipyretic
18	a_{18}	Antiseptic
19	a_{19}	Antispasmodic
20	a_{20}	Anti-tumor
21	a_{21}	Aphrodisiac
22	a_{22}	Astringent
23	a_{23}	Demulcent
24	a_{24}	Diaphoretic
25	a_{25}	Diuretic
26	a_{26}	Emmenagogue
27	a_{27}	Emollient
28	a_{28}	Hypotensive
29	a_{29}	Nutritive/Tonic
30	a_{30}	Relaxant

4.2 The Proximity Measures

The set of concepts is given as;

Concepts (C) = [C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}, C_{12}, C_{13}, C_{14}, C_{15}, C_{16}, C_{17}, C_{18}, C_{19}, C_{20}, C_{21}, C_{22}]

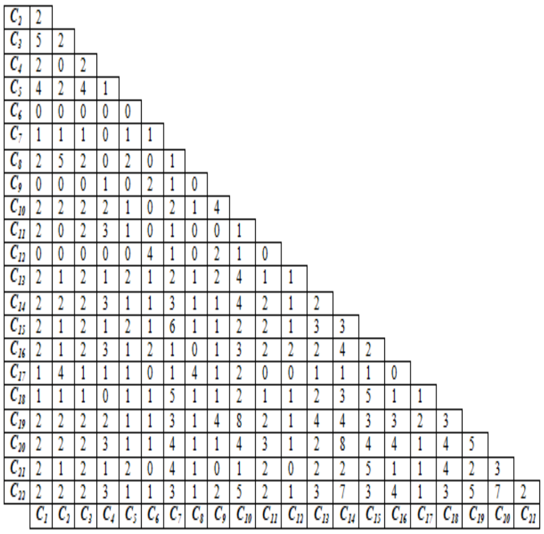
and the attributes as;

Attributes (A) = [a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}, a_{13}, a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, a_{19}, a_{20}, a_{21}, a_{22}, a_{23}, a_{24}, a_{25}, a_{26}, a_{27}, a_{28}, a_{29}, a_{30}].

Therefore, the multivariate set of concepts and attributes can be written as follows;

$C_1 = [a_3, a_5, a_7, a_{14}, a_{23}, a_{30}]$
$C_2 = [a_5, a_{14}, a_{16}, a_{17}, a_{18}, a_{27}]$
$C_3 = [a_3, a_5, a_{14}, a_{23}, a_{30}]$
$C_4 = [a_1, a_3, a_6, a_{12}, a_{19}, a_{30}]$
$C_5 = [a_5, a_{14}, a_{23}, a_{30}]$
$C_5 = [a_5, a_{14}, a_{23}, a_{30}]$
$C_6 = [a_8, a_{11}, a_{15}, a_{29}]$
$C_7 = [a_2, a_4, a_9, a_{14}, a_{22}, a_{29}]$
$C_8 = [a_5, a_{14}, a_{16}, a_{18}, a_{27}]$
$C_9 = [a_{11}, a_{12}, a_{20}, a_{25}, a_{26}, a_{29}]$
$C_{10} = [a_3, a_{12}, a_{14}, a_{17}, a_{20}, a_{21}, a_{25}, a_{29}]$
$C_{11} = [a_1, a_3, a_{22}, a_{30}]$
$C_{12} = [a_8, a_{11}, a_{15}, a_{28}, a_{29}]$
$C_{13} = [a_{14}, a_{21}, a_{25}, a_{29}, a_{30}]$
$C_{14} = [a_1, a_3, a_4, a_6, a_{10}, a_{13}, a_{14}, a_{17}, a_{29}]$
$C_{15} = [a_2, a_4, a_{14}, a_{22}, a_{29}, a_{30}]$

$C_{16} = [a_3, a_6, a_8, a_{17}, a_{29}, a_{30}]$
$C_{17} = [a_{12}, a_{14}, a_{16}, a_{18}, a_{27}]$
$C_{18} = [a_2, a_4, a_{14}, a_{22}, a_{29}]$
$C_{19} = [a_3, a_{12}, a_{14}, a_{17}, a_{20}, a_{21}, a_{22}, a_{25}, a_{29}]$
$C_{20} = [a_1, a_3, a_4, a_6, a_{10}, a_{14}, a_{17}, a_{22}, a_{29}]$
$C_{21} = [a_2, a_4, a_{14}, a_{22}, a_{30}]$
$C_{22} = [a_1, a_3, a_4, a_6, a_{14}, a_{17}, a_{25}, a_{29}]$



C_{2}	2																				
C_{3}	5	2																			
C_{4}	2	0	2																		
C_{5}	4	2	4	1																	
C_{6}	0	0	0	0	0																
C_{7}	1	1	1	0	1	1															
C_{8}	2	5	2	0	2	0	1														
C_{9}	0	0	0	1	0	2	1	0													
C_{10}	2	2	2	2	1	0	2	1	4												
C_{11}	2	0	2	3	1	0	1	0	0	1											
C_{12}	0	0	0	0	0	4	1	0	2	1	0										
C_{13}	2	1	2	1	2	1	2	1	2	4	1	1									
C_{14}	2	2	2	3	1	1	3	1	1	4	2	1	2								
C_{15}	2	1	2	1	2	1	6	1	1	2	2	1	3	3							
C_{16}	2	1	2	3	1	2	1	0	1	3	2	2	2	4	2						
C_{17}	1	4	1	1	1	0	1	4	1	2	0	0	1	1	1	0					
C_{18}	1	1	1	0	1	1	5	1	1	2	1	1	2	3	5	1	1				
C_{19}	2	2	2	2	1	1	3	1	4	8	2	1	4	4	3	3	2	3			
C_{20}	2	2	2	3	1	1	4	1	1	4	3	1	2	8	4	4	1	4	5		
C_{21}	2	1	2	1	2	0	4	1	0	1	2	0	2	2	5	1	1	4	2	3	
C_{22}	2	2	2	3	1	1	3	1	2	5	2	1	3	7	3	4	1	3	5	7	2
	C_{1}	C_{2}	C_{3}	C_{4}	C_{5}	C_{6}	C_{7}	C_{8}	C_{9}	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}	C_{20}	C_{21}

Figure 4.1: Triangular Matrix of Attributes Matches of the Domain Concepts

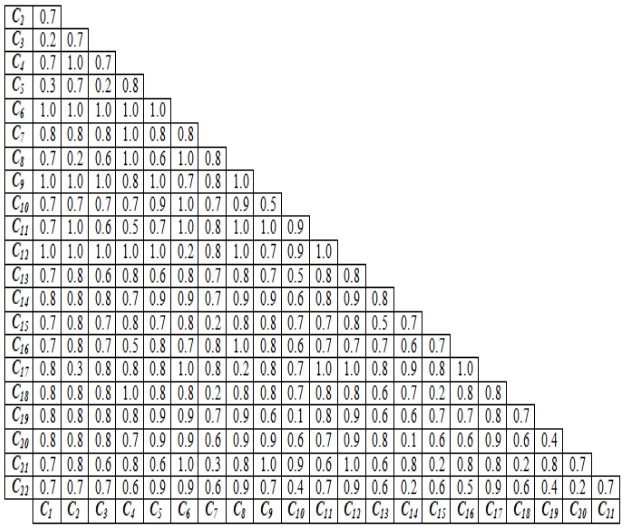
Afterward, attribute matches were derived from the set of multivariate data. In doing this, each concept in the array of concepts and attributes was compared with every other concept in the array, then the numbers of attribute matches were recorded in a triangular matrix. Figure 4.1 shows a triangular matrix of attributes matches of the concepts.

Furthermore, dissimilarity values were then computed as explained in section 3.2. However, in other to achieve scalability in computing the dissimilarity values, the highest numbers of possible attributes between two concepts whose dissimilarity are being computed were used as the denominator in the formula given in equation 1.4. Figure 4.2 shows the result of computing the dissimilarity values.

As shown in Figure 4.2, the dissimilarity values obtained ranged from 0.1 to 1.0. This shows that the concepts have varying relationships with one another in terms of their conceptual proximities. 0.1 implies very close proximity while 1.0 implies distant proximity. Locating concepts that are closely related to themselves was then carried out through the clustering algorithm presented in Section 3.3. The algorithm sections the concepts based on how close they are to one another.

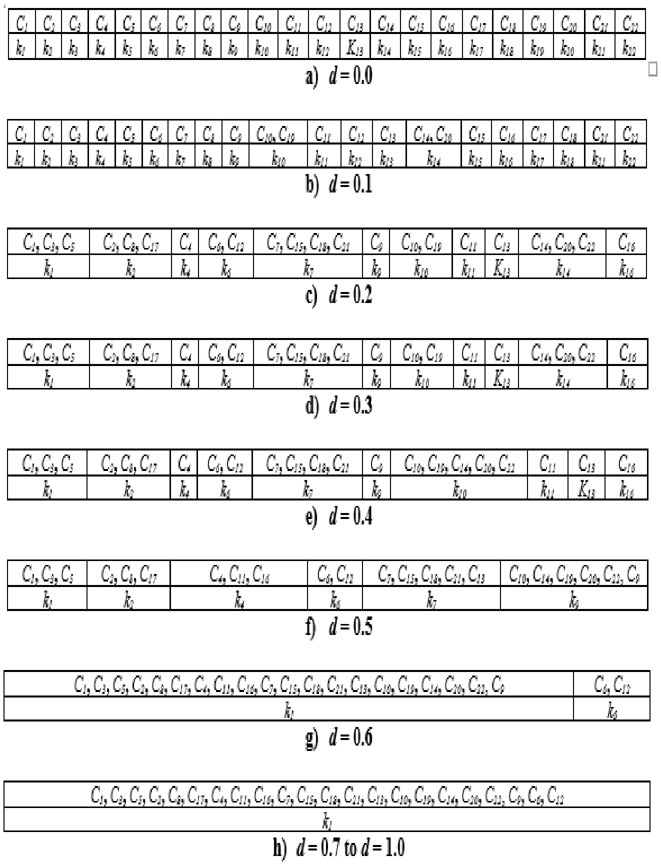
4.3 The Clustering Results

The clustering of concepts was carried out using the algorithm in section 3.3. Various iterative clusters of the dissimilarity values were generated in carrying this out. The initial value for the minimum dissimilarity was set at 0.0 (*d*) and an increment of 0.1 was added till the end of the iteration. Figure 4.3 shows the results of the concept clustering at each iterative stage.



C_{2}	0.7																				
C_{3}	0.2	0.7																			
C_{4}	0.7	1.0	0.7																		
C_{5}	0.3	0.7	0.2	0.8																	
C_{6}	1.0	1.0	1.0	1.0	1.0																
C_{7}	0.8	0.8	0.8	1.0	0.8	0.8															
C_{8}	0.7	0.2	0.6	1.0	0.6	1.0	0.8														
C_{9}	1.0	1.0	1.0	0.8	1.0	0.7	0.8	1.0													
C_{10}	0.7	0.7	0.7	0.7	0.9	1.0	0.7	0.9	0.5												
C_{11}	0.7	1.0	0.6	0.5	0.7	1.0	0.8	1.0	1.0	0.9											
C_{12}	1.0	1.0	1.0	1.0	1.0	0.2	0.8	1.0	0.7	0.9	1.0										
C_{13}	0.7	0.8	0.6	0.8	0.6	0.8	0.7	0.8	0.7	0.5	0.8	0.8									
C_{14}	0.8	0.8	0.8	0.7	0.9	0.9	0.7	0.9	0.9	0.6	0.8	0.9	0.8								
C_{15}	0.7	0.8	0.7	0.8	0.7	0.8	0.2	0.8	0.8	0.7	0.7	0.8	0.5	0.7							
C_{16}	0.7	0.8	0.7	0.5	0.8	0.7	0.8	1.0	0.8	0.6	0.7	0.7	0.7	0.6	0.7						
C_{17}	0.8	0.3	0.8	0.8	0.8	1.0	0.8	0.2	0.8	0.7	1.0	1.0	0.8	0.9	0.8	1.0					
C_{18}	0.8	0.8	0.8	1.0	0.8	0.8	0.2	0.8	0.8	0.7	0.8	0.8	0.6	0.7	0.2	0.8	0.8				
C_{19}	0.8	0.8	0.8	0.8	0.9	0.9	0.7	0.9	0.6	0.1	0.8	0.9	0.6	0.6	0.7	0.7	0.8	0.7			
C_{20}	0.8	0.8	0.8	0.7	0.9	0.9	0.6	0.9	0.9	0.6	0.7	0.9	0.8	0.1	0.6	0.6	0.9	0.6	0.4		
C_{21}	0.7	0.8	0.6	0.8	0.6	1.0	0.3	0.8	1.0	0.9	0.6	1.0	0.6	0.8	0.2	0.8	0.8	0.2	0.8	0.7	
C_{22}	0.7	0.7	0.7	0.6	0.9	0.9	0.6	0.9	0.7	0.4	0.7	0.9	0.6	0.2	0.6	0.5	0.9	0.6	0.4	0.2	0.7
	C_{1}	C_{2}	C_{3}	C_{4}	C_{5}	C_{6}	C_{7}	C_{8}	C_{9}	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}	C_{20}	C_{21}

Figure 4.2: Triangular Matrix of Dissimilarity Values of the Domain Concepts



C_{1}	C_{2}	C_{3}	C_{4}	C_{5}	C_{6}	C_{7}	C_{8}	C_{9}	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}	C_{20}	C_{21}	C_{22}
k_{1}	k_{2}	k_{3}	k_{4}	k_{5}	k_{6}	k_{7}	k_{8}	k_{9}	k_{10}	k_{11}	k_{12}	K_{13}	k_{14}	k_{15}	k_{16}	k_{17}	k_{18}	k_{19}	k_{20}	k_{21}	k_{22}

a) $d = 0.0$

C_{1}	C_{2}	C_{3}	C_{4}	C_{5}	C_{6}	C_{7}	C_{8}	C_{9}	C_{10}, C_{19}	C_{11}	C_{12}	C_{13}	C_{14}, C_{20}	C_{15}	C_{16}	C_{17}	C_{18}	C_{21}	C_{22}
k_{1}	k_{2}	k_{3}	k_{4}	k_{5}	k_{6}	k_{7}	k_{8}	k_{9}	k_{10}	k_{11}	k_{12}	k_{13}	k_{14}	k_{15}	k_{16}	k_{17}	k_{18}	k_{21}	k_{22}

b) $d = 0.1$

C_{1}, C_{3}, C_{5}	C_{2}, C_{8}, C_{17}	C_{4}	C_{6}, C_{12}	$C_{7}, C_{15}, C_{18}, C_{21}$	C_{9}	C_{10}, C_{19}	C_{11}	C_{13}	C_{14}, C_{20}, C_{22}	C_{16}
k_{1}	k_{2}	k_{4}	k_{6}	k_{7}	k_{9}	k_{10}	k_{11}	K_{13}	k_{14}	k_{16}

c) $d = 0.2$

C_{1}, C_{3}, C_{5}	C_{2}, C_{8}, C_{17}	C_{4}	C_{6}, C_{12}	$C_{7}, C_{15}, C_{18}, C_{21}$	C_{9}	C_{10}, C_{19}	C_{11}	C_{13}	C_{14}, C_{20}, C_{22}	C_{16}
k_{1}	k_{2}	k_{4}	k_{6}	k_{7}	k_{9}	k_{10}	k_{11}	K_{13}	k_{14}	k_{16}

d) $d = 0.3$

C_{1}, C_{3}, C_{5}	C_{2}, C_{8}, C_{17}	C_{4}	C_{6}, C_{12}	$C_{7}, C_{15}, C_{18}, C_{21}$	C_{9}	$C_{10}, C_{19}, C_{14}, C_{20}, C_{22}$	C_{11}	C_{13}	C_{16}
k_{1}	k_{2}	k_{4}	k_{6}	k_{7}	k_{9}	k_{10}	k_{11}	K_{13}	k_{16}

e) $d = 0.4$

C_{1}, C_{3}, C_{5}	C_{2}, C_{8}, C_{17}	C_{4}, C_{11}, C_{16}	C_{6}, C_{12}	$C_{7}, C_{15}, C_{18}, C_{21}, C_{13}$	$C_{10}, C_{14}, C_{19}, C_{20}, C_{22}, C_{9}$
k_{1}	k_{2}	k_{4}	k_{6}	k_{7}	k_{9}

f) $d = 0.5$

$C_{1}, C_{3}, C_{5}, C_{2}, C_{8}, C_{17}, C_{4}, C_{11}, C_{16}, C_{7}, C_{15}, C_{18}, C_{21}, C_{13}, C_{10}, C_{19}, C_{14}, C_{20}, C_{22}, C_{9}$	C_{6}, C_{12}
k_{1}	k_{6}

g) $d = 0.6$

$C_{1}, C_{3}, C_{5}, C_{2}, C_{8}, C_{17}, C_{4}, C_{11}, C_{16}, C_{7}, C_{15}, C_{18}, C_{21}, C_{13}, C_{10}, C_{19}, C_{14}, C_{20}, C_{22}, C_{9}, C_{6}, C_{12}$
k_{1}

h) $d = 0.7$ to $d = 1.0$

Figure 4.3: Diagram showing the Result of the Concept Clustering at each Iteration Stage

The clustering process resulted in eight (8) iteration stages as shown in Figure 4.3. The first iteration resulted into twenty-two (22) clusters, the second resulted in twenty (20) clusters, the third and fourth iterations both resulted into eleven (11) clusters, the fifth iteration resulted into ten (10) clusters, the sixth iteration resulted into six (6) clusters, the seventh iteration resulted into two (2) clusters, while the last iteration resulted into a single cluster. These imply that at the initial iteration, maximum homogeneity, and implicatively, minimum heterogeneity, is obtained.

This resulted in each of the concept forming its own cluster. The direct opposite of this is seen at the final iteration and there, minimum homogeneity and implicatively, maximum heterogeneity, is obtained. This resulted in all the concepts forming a single cluster. Therefore, obtaining the optimal homogeneity and heterogeneity is necessitated using cluster analysis and validation.

4.4 Cluster Analysis and Validation

The parameters considered for the analysis of the clusters are the number of clusters generated at each iteration point, the sum of all the mean of the entire clusters at each iteration point, the sum of all the within-cluster sum of squared errors, the total mean divided by the number of clusters generated at the iteration point and the total within-cluster sum of squared errors (WSS) divided by the number of clusters generated at the iteration point. Table 4.3 shows the results of the cluster analysis.

From the table, the number of clusters (*K*) generated diminishes with increasing dissimilarity value. This means that the heterogeneity increases with increasing the dissimilarity value which is a normal behaviour of clustering algorithms. Also, the total sum of all within-cluster sum of squared errors (*Total WSS*) increases with increasing dissimilarity value. This signifies an increase in the numbers of concepts per cluster which is also a normal behaviour of clustering algorithms. Finally, the total within-cluster sum of squared errors divided by the number of clusters (*Total WSS/K*) was found to also have a divergence at the fifth iteration (*d* = 0.4). This implies that the optimal homogeneity and heterogeneity of the concepts is at the fourth iteration (*d* = 0.3). This is shown in Figure 4.4.

Table 4.3: Cluster Analysis and Validation

S/N	*d*	*K*	*Total Mean*	*Total WSS*	*Total Mean/K*	*Total WSS/K*
1	0.0	22	0.00	0.00	0.00	0.00
2	0.1	20	0.10	0.02	0.01	0.00
3	0.2	11	0.66	0.17	0.06	0.02
4	0.3	11	0.66	0.17	0.06	0.02
5	0.4	10	0.77	0.26	0.08	0.03
6	0.5	6	1.21	0.58	0.20	0.10
7	0.6	2	0.63	0.60	0.32	0.30
8	0.7	1	0.55	0.62	0.55	0.62

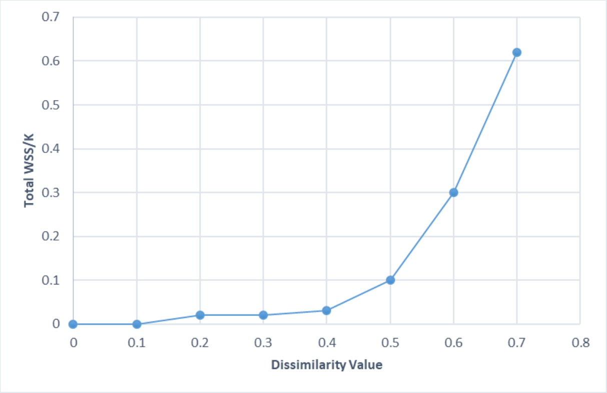


Figure 4.4: Graph Showing Total WSS/K by Dissimilarity Value

Inferentially, the optimal cluster of concepts is at the fourth iteration and it is given as k_1= $[C_1, C_3, C_5]$, k_2= $[C_2, C_8, C_{17}]$, k_4= $[C_4]$, k_6= $[C_6, C_{12}]$, k_7= $[C_7, C_{15}, C_{18}, C_{21}]$, k_9= $[C_9]$, k_{10}= $[C_{10,} C_{19}]$, k_{11}= $[C_{11}]$, k_{13}= $[C_{13}]$, k_{14}= $[C_{14}, C_{20}, C_{22}]$, k_{16}= $[C_{16}]$. This gives k_1= [Asthma, Cold & Catarrhal, Cough], k_2= [Chicken Pox, Eczema, Ring Worm], k_4= [Convulsions], k_6= [Diabetes, Hypertension], k_7= [Diarrheal, Pile, Stomach Disorder, Ulcer], k_9= [Fibroid], k_{10}= [Gonorrhea, Syphilis], k_{11}= [Hernia], k_{13}= [Erectile Dysfunction], k_{14}= [Malaria, Typhoid, Yellow Fever], k_{16}= [Rheumatism].

Implicatively, the concepts within the same cluster are more closely related to one another than to the other concepts in the domain. Consequently, domain activities and processes such as knowledge capture, knowledge representation, knowledge documentation and knowledge transfer of these concepts in the ATM practice should be carried out together as obtained in this research in order to obtain holistic and whole outcomes.

5.0 CONCLUSION

The multivariate clustering algorithm presented in this work provides a means of clustering multivariate data without the need for resolving the data into a univariate or bivariate data. More specifically, it was able to cluster the multivariate set of data which arose by attributing the medicinal properties of plants used in the treatment and/or management of some illnesses to the therapies of the illnesses. This provides a way of grouping both the medicinal plants as well as the illnesses. The grouping results in the modularization of the ATM ontology. Inferentially, the scalability, computability, and applicability of the ontology will be improved through the modularization.

REFERENCES

1. Sriharee, G. (2015). “An Ontology-Based Approach to Auto-Tagging Articles”. Vietnam Journal Computer Science Volume 2 pp 85–94 DOI:10.1007/s40595-014-0033-6

2. Krallinger, M., Leitner, F., Vazquez, M., Salgado, D., Marcelle, C., Tyers, M., Valencia, A. and Chatraryamontri, A. (2012). How to Link Ontologies and Protein-Protein Interactions to Literature: Text-Mining Approaches and the BioCreative Experience. *Database*, *2012*.

3. Bezerra C., Freitas F., Euzenat J., Zimmermann A. (2009). An Approach for Ontology Modularization. *Proc. Brazil/INRIA Colloquium on Computation: Cooperations, Advances and Challenges (Colibri)*, pp.184-189, 2009.

4. Atemezing, G., & Pavón, J. (2009). An ontology for African traditional medicine. In *International Symposium on Distributed Computing and Artificial Intelligence 2008 (DCAI 2008)*, pp. 329-337. Springer, Berlin, Heidelberg.

5. World Health Organization (2000). General Guidelines for Methodologies on Research and

Evaluation of Traditional Medicine. (2000). WHO/EDM/TRM/2000.1, pp. 1-74.

6. Paulo P. M., Kofi B. and Ossy M. K. (2010). Clinical Practices of African Traditional Medicine. *African Health Monitor*, 13, World Health Organization Africa. https://www.aho.afro.who.int/en/ahm/issue/13/reports/clinical-practices-african-traditional-medicine. Accessed on 10th May 2017.

7. Aletti, G. and Micheletti, A. (2017). A Clustering Algorithm for Multivariate Data Streams with Correlated Components. *J Big Data*. 4: 48. https://doi.org/10.1186/s40537-017-0109-0.

8. Wennerberg, P., Schulz, K., and Buitelaar, P. (2015). Ontology Modularization to Improve Semantic Medical Image Annotation. *Journal of Biomedical Informatics*, 44 (1), 155-162. 1532-0464/$ 2010 Elsevier Inc. All rights DOI: 10.1016/j.jbi.2010.12.005.

9. Soraya S. A., Mimoun M. and Sidi M. B. (2015). Ontology Partitioning: Clustering Based Approach. *I.J. Information Technology and Computer Science*, 2015, 06, 1-11 Published Online May 2015 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.20 15.06.01.

10. Konys, A. (2018). Towards Knowledge Handling in Ontology-Based Information Extraction Systems. *Procedia Computer Science*, 126, 2208-2218.

11. Awoniran O. M., Soriyan H. A., Elujoba A. A. and Okunnade A. A. (2019). Establishing Consensus in Knowledge Base Creation of Medicinal Plants of African Traditional Medicine Practice”. *Applications of ICT to Teaching Research and Administration (AICTTRA) Conference, 2019*.

12. Brennan, J., and Martin, E. (2012). Spatial proximity is more than just a distance measure International Journal of Human-Computer Studies, 70(1), 88-106.

13. Shai Shalev-Shwartz and Shai Ben-David (2014). “Understanding Machine Learning: From Theory to Algorithms”. Cambridge University Press New York, NY, USA ©2014 ISBN:1107057132 9781107057135.

INVESTIGATION OF FEATURE CHARACTERISTICS FOR YORUBA NAMED ENTITY RECOGNITION SYSTEM

*Adegunlehin E. A., Asahiah F. O. and Onifade M. T.
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.
Email of Corresponding Author: *adegunlehinabayomi@gmail.com

ABSTRACT
Researches on Named Entity recognition over the years has been concentrated on English and a few other European languages. However, Yoruba language and other African Languages have not received so much attention in this field. In this paper, we explored two features characteristics - Surrounding words and Part-of-Speech (PoS) tag of Named Entities as it can be applied to the recognition of Yoruba Named Entities. The paper considered mentions of only three named entities namely Person, Location and Time entities in a Yoruba text. The classification model was formulated using Conditional Random Fields (CRFs) classifier and a feature investigation experiment was carried on the result of the classification. The result of the experimental set up showed the impact of each of the features. PoS tag as a feature in Yoruba NER helps to identify a word as a Named Entity while information about the Surrounding words helps to classify a word into its Named Entity class. This paper has shown that to have a good Named Entity recognition system for Yoruba Text, choice of PoS and Surrounding Words as a feature set is crucial in the recognition of Named Entities in a Yoruba text.

Keywords: Named Entity Recognition, Yoruba, Conditional Random Field, Natural Language Processing.

1.0 INTRODUCTION

Information is largely on the increase publicly, however, one of the most abundant resources today is textual data. For this reason, information extraction (IE) emerged as a task to determine important entities, relations, and events in a human language text (Grishman, 2003). Name Entity (NE) is an expression that refers to proper names such as persons, locations, and organizations. Named Entity Recognition (NER) is a subtask of Human Language Processing applications. Usually NER is treated as a two-step process which is firstly the identification of proper names in text, and secondly the classification of these names into a set of predefined categories of interest, such as person names, organizations, etc (Chopra and Morwal, 2013).

Its usage can be found in Content recommendations in the media industry, Online journal or publication site, Customer Support, Legal and Medical Documents.

There are three basic approaches to NER: these are Rule-Based, Machine Learning based and Hybrid Approach. Rule-based approaches provide better results for restricted domains only, but they need immense experience and linguistic information of the specific language or domain (Shaalan and Raza, 2008**)**. Machine learning (ML) approaches are easily trainable and adaptable to different domains. But these techniques require large annotated corpus for training and testing (Benajiba *et al.*, 2009). Hybrid approaches take advantage of both rule-based and machine learning-based techniques. (Srikantha and Murthy, 2008).

Recognising entities in a text has been found to be key to many NLP applications, including Automatic Text Summarization (Nobata *et al*., 2003), Information Extraction Systems (Toda and Kataoka, 2005), Question-Answering Systems (Rodrigo et al. 2013), and Machine Translation (Steinberger *et al.*, 2011).

Yorùbá is regarded as one of the major well-spoken languages in Nigeria. The language is spoken in Nigeria and a few other African countries (Barber, 2015). The *Yorùbá* language (YL) is a tone language that belongs to the West Benue-Congo family of African languages and it is spoken by over forty million people (Fabunmi, 2005).

Several researchers have worked on building NER systems for different languages using various approaches. It is however impossible to use them directly for NER in Yoruba language due to various issues such as morphological differences and lack of digital resources. So, the development of various NLP applications for Yoruba languages is very dormant because many NLP applications are dependent on the availability and accuracy of NER systems.

In this paper, we seek to create an optimal set of features for the Yoruba Named Entity Recognition task. We explored two feature characteristics that can be used in performing Named Entity Recognition for Yoruba text. A comparison analysis was performed to show the impact of these features on the two-step process. This rest of the paper is structured as follows: Related works, Methodology, Result and Discussion, and Conclusion.

2.0 LITERATURE REVIEW

Ikechuckwu *et al.,* (2019) in their work titled "A First Step Towards the Development of *Yorùbá* Named Entity Recognition System", the authors focused on creating a ground for researchers on developing a robust NER for *Yorùbá* language. It involved the use of some widely used features such as PoS of a word, its surrounding words, affixation, capitalization, etc., and sequence modeling framework of other languages for *Yorùbá* in order to investigate their usefulness.

Sharnagat (2014) in his paper titled "Named Entity Recognition: A Literature Survey" explains different techniques being used for the NER task i.e. supervised techniques, semi-supervised and unsupervised techniques. The focus of this paper is to improve the NER system mainly for Indian languages.

Tkachenko and Simanovsky (2012) in their paper titled "Named Entity Recognition: Exploring Features" presented research on the complete features used in identifying a supervised-based NER task, various combinations of these features and evaluation of the performance. The work tries to reveal the effectiveness of clustering features and their combinations on NER.

A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi (Ekbal and Bandyopadhyay, 2009) describes CRF based NER. Prefix, suffix, tags of previous words, POS tags, first word, length of the word and gazetteer lists are some language inherited features described in this paper.

3.0 METHODOLOGY

This research uses the statistical-based ML technique in extracting and classifying NEs. Conditional Random Fields has been chosen as the ML algorithm.

3.1 Corpus

An annotated corpus is an essential resource for data-driven techniques. However, no publicly available NE annotated corpus exists for Yoruba that could be directly used. Therefore, for this research Yoruba corpus was created and manually annotated. The Named Entity categories considered are Person, Location and Time Entities. The corpus was annotated for PoS tags and Named Entity tags.

3.2 Recognition technique

Conditional Random Fields (CRFs) are undirected graphical models (Lafferty *et al.*, 2001) used to calculate the conditional probability of values on designated output nodes, given values on other designated input nodes (Wallach 2004). It has shown success in various sequence modeling tasks including NER tasks (Sha and Pereira 2003). Among CRF toolkits, CRF++ and CRFsuite are the most popular choices. However, CRFsuite which is more robust and faster to train was chosen for this work. It requires training and a testing data set. From the annotated corpus 80% of the data set is used as a train set and 20% used as a test set. Before the recognition process, the corpus has been prepared to align to the format required by the toolkit as shown in Table 1.1.

Table 1.1 Processed data

Sentence #	**Word**	**PoS**	**Tag**
Sentence 383	Ẹ̀gbọ́n	NN	O
	Ilẹ́sanmí	NNP	B-Per
	ní	VB	O
	Àdùnní	NNP	B-Per
	ìyá	NN	O
	Dúró	NNP	B-Per
	Orímóògùnjẹ́	NNP	I-Per
	.	.	O

3.4 Features Experiment

The purpose of this experiment is to evaluate the relevance of the above features in the recognition of each of the Named Entity categories identified in the data set. Following are the details of the set of features that have been applied to this experiment:

1. **Context words**: Preceding and next word of a particular word. This is based on the observation that the surrounding words are very effective in the identification of *Yorùbá* NEs. The context is defined as a window of +/- n tokens from the NE of interest, however, a context size of -1/+1 performs the best in this task. Some surrounding words serves as a keyword that sometimes co-occur with person names, location, and time expression. For example, titles and honorifics (*ọ̀gbẹ́ni, olóyè*) are the keywords for person names, while "*ọdún", "agogo", "ọ̀sẹ̀"* designators are for time expressions, and "*àdúgbò* ", "*ìlú*", etc are the keywords for locations. The keywords can help identify as well as classify the entities. Often times some NE do have some words which could serve as a pointer to the occurrence of such NE in a sentence. For instance words like "*ní* ", "*léhìn*", "*ní bí* ", "*fún*", "*ní*… NE… *séhìn*" etc, foretells a possible occurence of time expression. N-grams created from the words would serve as a feature for the classifier.
2. **Parts-of-Speech (POS) Information**: The POS of the current word and the surrounding words are used as a feature for NER. We realized that the detailed POS tagging is not very relevant i.e. having all the words in the corpus tagged. Since NEs are noun phrases, the noun tag is very relevant.

Therefore, to carry out the experiment, all possible feature combinations are used and evaluated. The result was used as a baseline for benchmarking the main experiment for this research. The standard evaluation metrics used are Precision, Recall, and F-measure. The baseline experiment in which all the feature combinations are used gave an F-measure value of 89.06%.

4.0 RESULT

Case 1 - Performance of the *Yorùbá* NER system without part of speech tags

In this case, PoS of tokens were not used as a feature but the other features were considered. The F-measure reduced by 7.04% from the baseline F-measure value. This result as shown in Table 4.1 shows that the use of the PoS tag of a word and PoS of surrounding words contributes to the recognition of the Named Entity tag of a *Yorùbá* word in a sentence.

Table 4.1 Performance of the *Yorùbá* NER system without Part-of-Speech tags

Metrics	Simple Average	Weighted Average
Recall	0.7704	0.7588
Precision	0.8955	0.8925
F-measure	0.8283	0.8202

Case 2 - Performance of the *Yorùbá* NER system without surrounding words

In this case, all features except the surrounding words were used. The surrounding words are the word preceding the present token and the next word after the present token. The performance of the system dropped greatly by 8.07% from the baseline F-measure value. As shown in Table 4.2, this result reveals that for *Yorùbá* Named Entity recognition task, words surrounding a particular word give a lot of information in predicting the Named Entity tag of that particular word.

Table 4.2 Performance of the *Yorùbá* NER system without surrounding words

Metrics	Simple Average	Weighted Average
Recall	0.8054	0.8099
Precision	0.8027	0.8058
F-measure	0.8041	0.8079

In summary, excluding PoS from the feature set, the system could not identify some words as a possible named entity, as a result, some named entity words are classified as non-entity. On the other hand, exclusion of Surrounding words from the feature set made the system not to correctly classify some words into their respective named entity class labels.

5.0 CONCLUSION

The focus of this paper was on the investigation of feature characteristics for Yoruba NE. We have proved that feature choice plays a crucial role in NE recognition. From the experiment conducted, each of the features used either helps to identify a word as a possible named entity or helps to classify such word to the proper named entity class it belongs. Due to the lack of linguistic resources like large NE annotated corpora and well-developed POS taggers, there are limitations entwined in this research. This paper concludes that large well-annotated training data and a well-developed PoS tagger would increase the overall accuracy of the system. Finally, one of the major and valuable outcomes of the research is the Yoruba NE annotated corpus which could be used directly for Yoruba NER related researches in the future.

REFERENCES

Barber, K. (2015). Yorùbá Language and Literature. In Oxford Bibliographies Online in African Studies. Retrieved July 12, 2019 from https:// www.oxfordbibliographies.com/view/document/obo-9780199846733/obo-9780199846733-0156.xml

Benajiba, Y., Diab, M., and Rosso, P. (2009). Arabic Named Entity Recognition: A Feature-Driven Study, IEEE Transactions on Audio, Speech, and Language Processing, 17(5), pp 926-934.

Chopra, D. and Morwal, S. (2013). Identification and Classification of Named Entities in Indian Languages. *International Journal on Natural Language Computing* (IJNLC). 2(1): 37-43.

Ekbal, A., Bandyopadhyay, S. (2009). A Conditional Random Field approach for named entity recognition in Bengali and Hindi, Germany: Department of Computational Linguistics, University of Heidelberg, India: Department of Computer Science and Engineering Jadavpur University.

Fabunmi, F. A., and Salawu, A. S. (2005). Is Yorùbá an Endangered Language? *Nordic Journal of African Studies*. 14(3): 391–408.

Grishman, R. and Sundheim, B. (1996), Message Understanding Conference-6: A Brief History. *In COLING*, 96, pp. 466-471.

Ikechukwu I, Adebayo O, and Bosede A. (2019). A First Step Towards the Development of Yorùbá Named Entity Recognition System. *International Journal of Computer Applications*. 182. 1-4.

Nobata, C. Sekine, S. and Isahara, H. (2003). Evaluation of Features for Sentence Extraction on Different Types of Corpora. *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*. http://aclweb.org/anthology/W03-1204. (Accessed: 2019-01-13).

Palmer, D. D. and Day, D. S. (1997). A Statistical Profile of the Named Entity Task. *In Proceedings of the Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Stroudsburg*, PA, USA, pp. 190-193.

Ratinov, L. and Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. In *Conference on Computational Natural Language Learning (CoNLL). Association for Computational Linguistics*. pages 147–155

Rodrigo, Á, Pérez-Iglesias, J., Peñas, A., Garrido, G. and Araujo, L. (2013). *Answering Questions About European Legislation. Expert Systems with Applications*, 40(15): 5811-5816

Sha, F. and Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. *In Conference on Human Language Technology and North*

American Association for Computational Linguistics (HLT-NAACL), pp. 213–220.

Shaalan, K., and Raza, H. (2008). Arabic Named Entity Recognition from Diverse Text Types, *In Advances in Natural Language Processing*, Springer Berlin Heidelberg. pp. 440-451.

Sharnagat, R. (2014). Named Entity Recognition: A Literature Survey, Bombay: Indian Institute of Technology.

Srikantha P. and Murthy K. N. (2008), Named Entity Recognition for Telugu. *In Proceedings* of IJCNLP-08 workshop on NER for South and Sound East Asian Languages. pp. 41-50.

Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., and Van der Goot, E., (2011). JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. *In Proceedings of the 2011 Conference on Recent Advances in Natural Language Processing*, pp. 104.

Tkachenko, M., and Simanovsky, A. (2012). Named Entity Recognition: Exploring Features. *In Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012)*, Vienna, Austria, pp 118-127.

Toda, H. and Kataoka, R. (2005). A search result clustering method using informatively named entities. *In Proceedings of the Seventh Annual ACM International Workshop on Web Information and Data Management, ACM*. pp. 81-86.

Wallach, H. M. (2004). Conditional random fields: An introduction, Technical Report MS-CIS-04-21, University of Pennsylvania, Department of Computer and Information Science, University of Pennsylvania.

AWARENESS, PERCEIVED USEFULNESS AND ACCEPTABILITY OF INSTITUTIONAL REPOSITORY BY LECTURERS IN OBAFEMI AWOLOWO UNIVERSITY, ILE-IFE, OSUN STATE, NIGERIA

Jamogha E., *Adeagbo O. O., Oguntuase F. Z. and Asubiojo B. O.
Hezekiah Oluwasanmi Library, Obafemi Awolowo University, Ile-Ife, Nigeria

*Email of Corresponding Author: *omoboladeadeagbo@oauife.edu.ng*

ABSTRACT
Institutional Repository (IR) is meant to archive digitally, the intellectual output of an institution of higher learning and it is dependent on the awareness and how acceptable it is for contributors to contribute their scholarly publications to the IR, hence, this study focused investigation on awareness, perceived usefulness and acceptability of IR by lecturers in Obafemi Awolowo University (OAU), Ile-Ife, Osun State, Nigeria. Descriptive survey was adopted for this study. Random sampling technique was used to select 140 respondents from a population of 1,399 lecturers in OAU. A total of 140 copies of questionnaire were distributed to respondents, however, 135 were duly filled and returned thereby, giving a response rate of 96.43%. Research question 1, 2, and 3 were answered using frequency and mean distribution, while question 4 was answered using frequency and simple percentage. The hypothesis was tested at a 0.05 level of significance, using regression analysis method. Findings of this study revealed that the awareness level of OAU IR by lecturers was low, as they were unaware of the IR's URL address or that IR can be accessed via the university's website. However, the perceived usefulness of IR, such as increasing publication citation and web metric ranking of the university was rated high. Similarly, acceptability of the IR, such as accepting to add research publications to and made accessible to all via the IR had a high rating. The perceived usefulness and awareness of IR have joint positive influence on the acceptability of IR. The challenges hindering IR acceptability were lack of training and lukewarm attitude to IR by academics. Thus, it was recommended that awareness programmes and regular training on IR use should be organised for academics.

Keywords: Institutional repository, Institutional repository awareness, Institutional repository perceived usefulness, Institutional repository acceptability

1.0 INTRODUCTION

Intellectual outputs by institutions of higher learning are to be harnessed for the good of the institution in particular and the world in general, hence, there is agitation by academic communities to expose to the world, their scholarly works, such as research papers, course materials, lecture notes and other relevant information contents. However, they would have to be aware and be ready to accept and manage any platform that is useful for the management of scholarly works and making them readily available and accessible to as many that are interested. To this end, Institutional Repository (IR) provides a system for managing the intellectual output in an academic community. This was clearly observed by Twara and Gandotra (2018) that IR is now a platform for academicians and researchers worldwide that help, to break the barrier of time and space in accessing scholarly information resources.

Institutional Repository was described by Prabhakar and Rani (2017) as a term for the collection, management, preservation and the dissemination of scholarly works created in digital form by staff and students in individual institutions of higher learning. The creation of Institutional Repository (IR) by higher institutions of learning is aimed at developing a systematic and integrated approach to identifying, capturing, preserving and retrieving institutions' scholarly outputs such as, research papers, course materials, lectures note, and other relevant information content of the institution. IR relies on inputs from members of the academic community, hence scholars and researchers ought to have awareness of IR, so that they can contribute their quotas to the creation and the growth of the IR.

Beyond the awareness of IR by academics, IR's acceptability is the key to exploring the gains that IR presents to its user. The Cambridge Dictionary (2019) defines acceptability as the quality of being satisfactory and being able to agree to, and approved of. Therefore, acceptability of IR is an indication that it has a satisfactory quality that users would appreciate and are ready to agree with, in terms of conditions and willingly to approve it for use by others. However, for a potential user to accept IR as a platform for managing his/her scholarly publications and being able to access those of other scholars, he/she would have conceptualized the perceived usefulness of the IR. The usefulness of IR according to Ukwoma and Dike (2017) is that, it archives research publications to increase citation and visibility and the preservation of scholarly communications. Perceived usefulness was defined by Davis (1989) as "the degree to which a person believes that using a particular system would enhance his/her job performance". Therefore, perceived usefulness of IR in this study would be the extent to which users believe that using institutional repository would enhance their intellectual productivity.

This study is targeted at lecturers at the Obafemi Awolowo University (OAU), Ile-Ife, Osun State, Nigeria. The lecturers constitute a greater part of the population that is expected to contribute their scholarly publications to the pool of OAU institutional repository. They contribute directly through granting permission to add their publications to the IR or indirectly through supervision of other researchers that may likely add their intellectual output to the IR. Hence, this study focused investigation on the awareness, perceived usefulness and acceptability of IR by lecturers in OAU.

1.1 OAU Institutional Repository

The OAU institutional repository is the archive for the capturing, preserving and retrieving of the intellectual outputs of the institution. The OAU IR employs the use of DSpace software as the platform for the management of the institution scholarly output. The IR is centrally managed by librarians at the Hezekiah Oluwasanmi Library, OAU. The IR is Internet based, providing unrestricted access to its content via a Uniform Resource Locator (URL) address: “ir.oauife.edu.ng”. The IR can also be accessed via OAU website “oauife.edu.ng”.

1.2 Statement of the Problem

Institutional repository (IR) is meant to archive digitally, the intellectual output of an institution of higher learning, like the Obafemi Awolowo University. IR is expected to be populated with the institution’s scholarly publications such as research articles, inaugural lectures, theses and dissertation. The population and growth of the IR would depend on the awareness and how acceptable it is for the contributors to contribute their scholarly publications to the IR. The major contributors to the IR are the lecturers who will have to permit the addition of their research publications to the IR, thereby providing exposure and unrestricted access. However, observation has shown that the level of acceptability of IR by lecturer is generally low. This could be as a result of the level of awareness and the perceived usefulness of the IR. Hence, this study focused investigation on the awareness, perceived usefulness and acceptability of IR by lecturers in OAU.

1.3 Research Objectives

The main objective of this study is to investigate the awareness, perceived usefulness and acceptability of IR by lecturers in OAU. While the specific objectives are to:

(i) Investigate the awareness level of IR by lecturers in OAU.
(ii) Examine the perceived usefulness of the IR by lecturers in OAU.
(iii) Determine the extent of acceptability of IR by lectures in OAU.
(iv) Investigate the challenges hindering the acceptability of IR by lecturers in OAU.
(v) Determine the joint effect of awareness and perceived usefulness on the acceptability of IR by lecturers in OAU.

1.4 RESEARCH QUESTION

(i) What is the awareness level of IR by lecturers in OAU?
(ii) What is the perceived usefulness of IR by lecturers in OAU?
(iii) What is the extent of acceptability of IR by lectures in OAU?
(iv) What are the challenges hindering the acceptability of IR by lecturers in OAU?

1.5 Research Hypothesis

There is no significant joint influence of awareness and perceived usefulness on the acceptability of IR by lecturers in OAU.

2.0 REVIEW OF RELATED LITERATURE

The awareness of Institutional Repository (IR) has gained momentum the world over; this was revealed in Cullen and Nagata (2008) observation that, the Ministry of Education, Culture, Sports, Science and Technology in Japan, had encouraged university libraries in Japan to develop IR to promote knowledge sharing throughout Japan and internationally. Bamigbola and Adetimirin (2017) found that lecturers were aware of IRs. Francke, Gamalielsson and Lundell (2017) observed that since the early 2000s, IR has become a known infrastructure in institutions of higher education worldwide. Similarly, Vrana (2010) opined that IR is now being recognized as essential infrastructure for scholarly publishing in digital world.

Anunobi and Okoye (2008) noted that, IR provides a way of reducing the cost of scholarly publication and increasing the visibility and access to scholarly research. Musa, Umar and Abdulkadir (2014) examined “institutional digital repositories in Nigerian: issues and challenges”, they found that the increase of Nigerian content on the Web via institutional repository will enable meaningful and fruitful researches to be conducted through the access to relevant data and literatures. Mgonzo and Yonah (2014) found that IR is a viable solution to the major challenges encountered in managing and sharing scholarly information in institutions of higher learning. Ranasinghe and Chung (2018) highlighted the usefulness of IR by resolving that, it serves as a platform for scholarly publishing, and enable institutions to take control over their research output while promoting accessibility to the institutions research in an open-access environment.

A study conducted by Ukwoma and Dike (2017) to investigate the “academics’ attitude toward the utilization of institutional repositories in Nigerian Universities” shows that the academics in Nigerian universities have positive attitude towards the use of IR and are willing to submit their publications. This implies that there is a positive level of acceptability among the academics, given that they are willing to contribute their publications to the IR.

In Kenya, Moseti (2016) investigated “digital preservation and institutional repositories: case study of universities of Kenya”, and found that scholars in these universities personally engaged in preserving their digital collection but were not extensively using these universities institutional repositories, this was attributed largely to lack of awareness on the usefulness of IR to digital preservation. Hence, it was suggested that more awareness needed to be done. Ogbomo and Muokebe (2015) observed that the success of IR depends on the awareness of and the perceived usefulness of IR by lecturers.

Pickton and Barwick (2006) identified the challenges of IR to include: difficulties in generating content, issues such as, right to IR management, working culture, policy, lack of incentives and recurrent cost. Prabhakar and Rani’s (2017) study on “benefits and perspectives of institutional repositories in academic libraries” revealed that, formulating collection policy, bibliographic details standardization, checking copyright policy, and persuasion of lecturers to deposit their publications were challenges facing IR. Similarly, Faleh, Hani and Khaled (2011) found that acceptability, collection building, infrastructure, access restriction, authentication, copyright, strategic issues, funding and user interface were IR challenges. Adabayo (2013) noted that poor contributors support, tracking of publications, deterioration of media, technological obsolescence and funding were challenges to IR acceptability.

The literature so far reviewed shows that IR is useful for archiving scholarly output of an institution, however, there is low level of awareness among persons that are supposed to contribute their publications to the IR, this was evident in some of the literature that identified the challenge of IR acceptability and the difficulty of persuading lecturers to deposit their publications in the IR for archiving, visibility and accessibility. It is noteworthy that, none of the literature reviewed was on awareness, perceived usefulness and acceptability of IR by lecturers in OAU. It is this gap that this study intends to fill.

3.0 METHODOLOGY OF THE STUDY

Descriptive survey was adopted for this study. Random sampling technique was used to select 140 respondents from a population of 1,399 lecturers across the faculties in Obafemi Awolowo University. The instrument for data collection was questionnaire. The research questions were analysed using descriptive statistics, while inferential statistic was used to test the hypothesis at a 0.05 level of significance. One hundred and thirty-five copies of the questionnaire were used for data analysis.

Questionnaire administration and response rate

Response rates to the questionnaire is presented in Table 1

Table 1: Lecturers’ questionnaire distribution and response rate

No. Administered	No. Returned
140	13
100%	96.43%
N= 135	

N= Count

Table 1 shows that the total number of questionnaire administered to the respondents were one hundred and forty (140) copies. However, only 135 copies of the questionnaire were duly filled and returned with useful responses by the respondents giving 96.43% response rate.

2.1 Demography distribution of the undergraduates

This section focuses on the demographic characteristics of the respondents. Table 2 presents information on the demographic data of the respondents.

Table 2: Demographic data of respondents

Demography		Freq-uency	Percent-age
Sex	Male	109	80.70%
	Female	26	19.30%
Age	20-30	2	1.50%
	31-40	43	31.90%
	41-50	64	47.40%
	51-60	17	12.60%
	61-70	9	6.70%
Highest qualification	BSc	3	2.20%
	Masters	66	48.90%
	MPhil	4	3.00%
	PhD	62	45.90%
Status	Graduate Assistant	3	2.20%
	Assistant Lecturer	31	23.00%
	Lecturer II	36	26.70%
	Lecturer I	38	28.10%
	Senior Lecturer	16	11.90%
	Reader	7	5.20%
	Professor	4	3.00%
N=135			

N=Count

Results in Table 2 shows that most of the respondents 109(80.7%) were male, while 26(19.3%) were female. This implies that the male lecturers in OAU may likely have more influence on the acceptability of OAU’s IR when compared to their female counterparts. Table 2 also shows that majority of the respondents 64(47.4%) were within the age bracket of 41-50 years. The highest academic qualification of most of the respondents were Masters 66(48.9%) followed by PhD 62(45.9%). The status of most of the respondents 38(28.1%) were Lecturer I followed by Lecturer II 36(26.7%). This implies that most of the respondents would be actively involved in research as it is a prerequisite for the advancement of their career, and consequently, leading to an increase in scholarly publications that would constitute a substantial part of the OAU Institutional Repository (IR).

Research question 1: What is the awareness level of IR by lecturers in OAU?

The awareness level of IR by lecturers in OAU is presented in Table 3.

Table 3: Awareness level of IR by lecturers in OAU

No.	Statements	NA	FA	A	VMA	Mean
1	I am aware that members of staff can contribute their research publications to the IR	31	16	35	53	2.81
2	I am aware; the growth of IR is dependent on my contributions to its contents	47	9	28	51	2.61
3	I am aware that IR is open for all to access	31	16	66	22	2.59
4	I am aware that I can add my research work to the IR	31	44	15	45	2.55
5	I am aware that IR exist in OAU	31	32	49	23	2.47
6	I know that IR would provide clue to the university researches housed in the library	47	11	43	34	2.47
7	I am aware that IR would increase the visibility of my research publications	31	33	49	22	2.46
8	I am aware that PG students can submit their theses to the OAU IR	47	17	33	38	2.46
9	I know that the OAU IR can be accessed via the Internet	31	35	46	23	2.45
10	I am aware that I can access the OAU IR, via the university website	30	41	50	14	2.36
11	I am aware that my paper publications can be digitally converted and added to the IR	47	11	70	7	2.27
12	I am aware that the OAU IR can be accessed from any location in the world	45	36	44	10	2.14
13	I know the URL address of OAU IR	58	28	35	14	2.04
	N= 135, Calculated mean = 2.44, Criterion mean= 2.5					

N= Count, NA – Not Aware, FA – Fairly Aware,
A- Aware, VMA- Very Much Aware

Table 3 shows that most of the lecturers are aware that: members of staff can contribute their research publications to the IR (mean= 2.81), the growth of IR is dependent on their contributions to its contents (mean= 2.16), IR is open for all to access (mean= 2.59) and that they can add their research work to the IR (mean= 2.55). However, most of the lecturers (mean= 2.04) do not know the URL address of OAU IR, they are unaware that: OAU IR can be accessed from any location in the world (mean= 2.14), their paper publications can be digitally converted and added to the IR (mean= 2.27) and they can access the OAU IR, via the university website (mean= 2.36). From Table 3, the calculated mean of 2.44 is lesser than the criterion mean of 2.5; this implies the awareness level of IR by the lecturers in OAU is generally low. For instance, the URL address of OAU IR and the use of OAU's website that would facilitate access to the IR have a low awareness level.

Research question 2: What is the perceived usefulness of the IR by lecturers in OAU?

The perceived usefulness of IR by lecturers in OAU is presented in Table 3.

Table 4: Perceived usefulness of IR by lecturers in OAU

No.	Statements	SD	D	A	SA	Mean
1	IR can help to increase the citation of my research publications	0	0	39	96	3.71
2	IR can make access to scholarly publications easier	5	5	54	71	3.41
3	IR can enhance scholarly communication among academic staff	3	2	76	54	3.34
4	IR can improve the web metric ranking of the university	0	0	106	29	3.21
5	IR is an effective backup for the intellectual output of the university	15	6	53	61	3.19
6	IR is an effective medium for promoting scholarly publishing	1	2	104	28	3.18
7	IR can improve my scholarly impact on other researchers	9	16	62	48	3.1
8	IR can improve the visibility of my research publications	10	6	91	28	3.01
9	IR can provide a platform for e-learning	20	20	61	34	2.81
10	IR is effective for the digitization of research papers	20	7	93	15	2.76
	N=135, Calculated mean= 3.17, Criterion mean= 2.5					

N= Count, SD- Strongly Disagree, D- Disagree,
A- Agree, SA- Strongly Agree

Table 4 shows the perceived usefulness of IR, as most of the respondents indicated that “IR can help to increase the citation of my research publications” (mean= 3.17), “IR can make access to scholarly publications easier” (mean= 3.41), “IR can enhance scholarly communication among academic staff” (mean= 3.34), “IR can improve the web metric ranking of the university” (mean= 3.21) and “IR is effective for the digitization of research papers” (mean= 2.76). From Table 4, the calculated mean of 3.17 is greater than the criterion mean of 2.5. This implies that the lecturers generally have a positive perception that the OAU IR is useful.

Research question 3: What is the extent of acceptability of IR by lectures in OAU?

The extent of IR acceptability by lecturers in OAU is presented in Table 3.

Table 5: Extent of acceptability of IR by lecturers in OAU

No.	Statements	NA	FA	A	HA	Mean
1	I would recommend that PG theses be added to the IR	0	20	34	81	3.45
2	I will accept that my research work be added to the IR	0	20	55	60	3.3
3	I would recommend that the development of IR be adequately supported by the university management	0	20	59	56	3.27
4	I would accept that my paper publications be converted into digital format and made available on the IR	0	20	60	55	3.26
5	I would accept to personally add my research work to the IR	18	2	51	64	3.19
6	I would accept to create awareness of IR to my colleagues	0	20	70	45	3.19
7	I can recommend that IR be used to preserve the universities publications	0	20	77	38	3.13
8	I would accept that theses I supervised be added to the OAU IR	20	0	58	57	3.13
9	I will join in advocating for the embrace of IR by academics	0	31	69	35	3.03
10	I would accept that my research publications be made accessible to all via the IR	18	2	86	29	2.93

N=135, Calculated mean=3.19, Criterion mean= 2.5

NA- Not Accepted, FA- Fairly Accepted,
A- Accepted, HA- Highly Accepted

Table 5 shows the acceptability level of IR, as most of the respondents indicated that “I would recommend that PG theses be added to the IR” (mean= 3.45), “I will accept that my research work be added to the IR” (mean= 3.30), “I would recommend that the development of IR be adequately supported by the university management” (mean= 3.27) and “I would accept that my research publications be made accessible to all via the IR” (mean= 2.93). On a general note, since the calculated mean of 3.19 is greater than the criterion mean of 2.5 in Table 5, it implies that the acceptability level of OAU’s IR by lecturers is high.

Research question 4: What are the challenges hindering the acceptability of IR by lecturers in OAU?

The challenges hindering the acceptability of IR by lecturers in OAU, is presented in Table 6.

Table 6: Challenges hindering the acceptability of IR by lecturers in OAU

		Disagree		**Agree**	
No.	**Statements**	**N**	**Percentage**	**N**	**Percentage**
1	Users are rarely trained to use the IR	7	5.20%	128	94.80%
2	Inadequate sensitization on the benefits of IR	24	17.80%	111	82.20%
3	Lukewarm attitude to IR by academics	50	37.00%	85	63.00%
4	Lack of institutional support	64	47.40%	71	52.60%
5	Indifference to contributing to the IR	65	48.10%	70	51.90%
6	There is difficulty in accessing the IR	71	52.60%	64	47.40%
7	I lack the skills needed to use the IR	86	63.70%	49	36.30%

N= 135

N=Count

Table 6 shows the challenges hindering the acceptability of IR by lecturers, as most of the lecturers agreed that “users are rarely trained to use the IR” 128 (94.8%), followed by “Inadequate sensitization on the benefits of IR” 111 (82.2%), “lukewarm attitude to IR by academics” 85 (63.0%) and “Indifference to contributing to the IR” 70(51.9%). While most of the lecturers disagreed to statement such as “I lack the skills needed to use the IR” 49 (36.3%) and “There was difficulty in accessing the IR” 64 (47.4%). Therefore the challenges hindering the acceptability of IR are: lack of training, inadequate sensitization on the benefits of IR and the lukewarm attitude to IR by academics.

Research Hypothesis

There is no significant joint influence of awareness and perceived usefulness on the acceptability of IR by lecturers in OAU.

Table 7: Correlation showing joint influence of awareness and perceived usefulness on acceptability of IR by lecturers in OAU

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.835[a]	0.697	0.692	3.454

ANOVA[a]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	3616.045	2	1808.023	151.547	.000[b]
	Residual	1574.814	132	11.93		
	Total	5190.859	134			

a. Dependent Variable: IR Acceptability

b. Predictors: (Constant), IR Perceived Usefulness, IR Awareness

Table 7 reveals that coefficient of determination R^2 = .697, gives 69.7% proportion of variance. This implies that the independent variables (IR awareness and IR perceived usefulness) accounted for 69.7% of the variance in the dependent variable (IR acceptability). Hence, the joint influence IR awareness and IR perceived usefulness on IR acceptability (*F-ratio* = 151.547; *P(.000)*< 0.05). This means that 69.7% of the variance was accounted for by the predictor variables when taken together. Any other variable not included may have accounted for the remaining variance. Therefore, there is significant joint influence of IR awareness and IR perceived usefulness on IR acceptability by lecturers in OAU. Hence, the null hypothesis is rejected. Consequently, an improvement in IR awareness and perceived usefulness would result in an improvement in IR acceptability.

4.0 DISCUSSION OF THE FINDINGS

Findings of the study showed that the awareness level of IR by the lecturers in Obafemi Awolowo University (OAU) is generally low. For instances, the URL address of OAU IR and the use of OAU's website that would facilitate access to the IR have a low awareness level. This negates the observation of Francke, Gamalielsson and Lundell (2017) that IR has become a known infrastructure in institutions of higher education worldwide.

Secondly, it was found that IR was perceived to be highly useful. For instance, most of the respondents believe that IR can help to increase the citation of their research publications, increase the web metric ranking of the university, and it is effective for the digitization of their research papers. This is corroborated by the findings of Ukwoma and Dike (2017) and Anunobi and Okoye (2008) who noted that, IR provides a way of reducing the cost of scholarly publication and increasing citation, visibility and access to scholarly research.

Thirdly, it was found that the acceptability level of IR was high. For instance, the respondents accepted that they would recommend that postgraduate theses be added to the IR, and are ready to add their research publication to the IR and they also accept that their research publications can be made accessible to all via the OAU IR. This is affirmed by the Ukwoma and Dike (2017) who found that academics in Nigerian universities are willing to submit their publications for addition to the IR.

Fourthly, challenge hindering the acceptability of IR by lecturers, include lack of training, inadequate sensitization on the benefits of IR and the lukewarm attitude to IR by academics and indifference to contributing to the IR. This is affirmed by Prabhakar and Rani's (2017) who found that persuading of lecturers to deposit their publications was challenges facing IR. However, this negates the findings of Ukwoma and Dike (2017) who found that academics in Nigerian universities have positive attitude towards the use of IR.

Lastly, the study revealed that IR awareness and IR perceived usefulness had a joint influence on IR acceptability by lecturers in OAU. This implies that an improvement in IR awareness and perceived usefulness will result in an improvement in the acceptability of IR. This is affirmed by Ogbomo and Muokebe (2015) observation that the success of IR depends on the awareness of, and the perceived usefulness of IR by lecturers. For instance, the findings of Moseti (2016) that scholars in universities in Kenya were not extensively using their universities' institutional repositories, was attributed to lack awareness on the usefulness of IR to digital preservation, even when they personally preserve their publications digitally.

5.0 CONCLUSIONAND RECOMMENDATION

The acceptability of Institutional Repository (IR) by lecturers in OAU would be enhanced if there is a high level IR awareness and if the perceived usefulness of IR is factored into the provision and running of IR in OAU.

In line with the findings and conclusion made the following are therefore recommended:

(i) Since IR awareness level is low, more awareness programs and training on IR use should be organized for academics.

(ii) The perceived usefulness of IR was positive, hence, there is need for it to be sustained by library management, by ensuring that IR policies and designs are skewed towards making the OAU's IR useful to the users so that its' benefits can be maximized.

(iii) The acceptability of IR was positive, hence, it should be sustained by ensuring that the IR design is user friendly and all facility needed to support and assist the users should be provided, and awareness programmes and regularly training on IR use should be organized academics.

(iv) Since IR awareness and perceived usefulness has positive influence on the acceptability of IR by lecturers, all stakeholders should help in promoting the use of IR and its benefits whenever the opportunity shows forth.

Recommendation 1 to 4 can also be applied to address the issue of indifference by lecturers in contributing to the IR.

REFERENCES

Adebayo, E. (2013). Digital institutional repository: the challenges of instituting a viable library policy in Nigerian universities libraries: *a lead paper presented at the Annual general Meeting of The Nigerian Library Association, Ogun State Chapter, held on Thursday, 17th January, 2013 at the Neuro-Psychiatric Hospital, Aro, Abeokuta*.

Anunobi, C.V. & Okoye, I.B. (2018). The role of academic libraries in universal access to print and electronic resources in the developing countries, *Library Philosophy and Practice*

Bamigbola, A.A. & Adetimirin, A.E. (2017). Evaluating use of institutional repositories by lecturers in Nigerian universities.*Information Impact: Journal of Information and knowledge management, 8*(3): 83-102.

Barwick, J. & Pickton, M. (2006). A librarian's guide to institutional repositories. *eLucidate, 3* (2): 3-9

Cambridge Dictionary (2019). Acceptability. United Kingdom: Cambridge University Press. Retreived from: www.google.com/amp/s/dictionary.combridge.org/amp/english/acceptability?espv=1

Cullen, R.. & Nagata, H. (2008). Academic libraries in Japan. *Journal of Academic Librarianship, 34* (2), 163-167.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and users acceptance of information technology. *MIS Quarterly, 13*(3): 319-340.

Faleh, A.A., Hani, J.I. & Khaled, B.H. (2011). Building a knowledge repository: linking Jordanian Universities e-library in an integrated database system.

Francke, H., Gamalielsson, J. & Lundell, B. (2017). Institutional repositories as infrastructures for long-term preservation. *Information research, 22*(2), paper 757. Retrieved from http://InformationR.net/ir/22-2/paper757.

Mgonzo, W.J. & Yonah, Z.O. (4014). Design and development of web based digital repository for scholarly communication: a case of NM-AIST Tanzania. *International Journal of Knowledge Content Development & Technology, 4*(2): 97-108

Moseti, I. (2016). Digital preservation and institutional repositories: case study of Universities of Kenya. *Journal of South African Society of Archivist, 49*: 137-254

Musa, A.U., Umar. S, & Abdulkadir, A. (2014). Institutional digital repositories in Nigerian: issues and challenges. *IOSR Journal of Humanities and Social Science (IOSR-JHSS), 19*(1):16-21

Obgomo, E.F. & Muokebe, B.O. (2015). Institutional repositories as emerging initiative in Nigerian university libraries. *Information and Knowledge Management, 5*(1): 1-8

Prabhakar, S.V.R. & Rani, S.V.M. (2017). Benefits and perspectives of institutional repositories in academic libraries. *Scholarly Research Journal for Humanity Science & English Language.* Online ISSN 2348-3083, SJ IMPACT FACTOR 2016 = 4.44, www.srjis.com UGC Approved Sr. No.48612, DEC-JAN 2018, VOL- 5/25

Tiwari, S. & Gandotra, N. (2018). Fundamental of concepts of institutional repositories. *Journal of Advancements in Library Sciences, 5* (2): 51-54.

Ranasinghe, W,M,T. & Chung, J.M. (2018). Institutional repository based open access scholarly publishing system: a conceptual model. *Library Philosophy and Practice (e-journal),* http://digitalcommons.unl.edu/libphilprac/2014

Ukwoma, S.C. & Dike, V,W. (2017). Academics' attitude toward the utilization of institutional repositories in Nigerian Universities. *Portal:Library and the Academic, 17* (1): 17-32

Vrana, R. (2010). Digital repositories and the future of preservation and use of scientific knowledge. *Informatol, 44*(1): 55-62.

A CONCEPTUAL MODEL FOR GRAPH-BASED KEYPHRASE EXTRACTION IN DOCUMENTS

***[1]Abimbola R. O., [2]Awoyelu I. O. and [3]Hunsu F. O.**

[1, 2] Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.

[3] Department of English, Obafemi Awolowo University, Ile-Ife, Nigeria.

*Email of Corresponding Author: *rilwanabimbola@yahoo.com*

ABSTRACT

The study presents a conceptual model for graph-based keyphrase extraction in documents. Keyphrase extraction is a method that is used to extract important phrases that are representative of a document. A graph-based approach models a document as a graph using nodes and edges to represent the phrases and relationships between the phrases respectively. Existing graph-based approaches that used text-rank algorithm for keyphrase extraction lack a good clustering method in extracting keyphrases. Also, extracted keyphrases do not contain nouns as part of them. This paper focusses on extraction of keyphrases that are noun-centric. A graph-based model for keyphrases extraction was formulated using text-rank algorithm, affinity propagation, k-means algorithm, cosine similarity and noun phrase identifier. The formulated model was designed using Unified Modelling Language. The work proposes an algorithm that can be implemented for keyphrase extraction.

Keywords: Keyphrases, Keywords, Graph-based, Text Rank, Noun-centric.

1.0 INTRODUCTION

Conceptual modelling is known to be an essential step in simulation to provide a better understanding of the system in early stages of development, and in this way to increase quality of requirements, enhance communication between users and developers and help in verification and validation of software products (Bozlu and Demirörs, 2008). Additionally, conceptual model acts as a basis to judge interoperability among distributed simulation systems.

Graph model represents a real-life abstraction that uses the graph theory to denote an abstraction with edges and nodes that allows future investigation on the real-life abstraction. This method constructs a word graph based on word co-occurrences within the document first and it ranks the words according to their scores. As a result, the top ranked words are the important key words. In using the graph model for keyphrase extraction, text units are first identified and added as nodes in a graph. Then, meaningful relations between the text units are identified and are added as edges between the nodes. The resulting graph is used to compute the importance of each node in relation to all other nodes. Similarly, the score of a particular node will affect the scores of other nodes connected to it.

Keyphrase extraction is the method that identifies group of words that can provide more description about a particular document. Keyphrases are used to depict words or group of words that are used to represent important information about a document. Keyphrase extraction has been a technique that is done by human annotators that select phrases from a list of vocabulary or assign what they feel are keyphrases of the documents at hand (El-Bethagy and Rafea, 2009). In the case of selecting phrases from a list of vocabulary, the phrases selected may not appear in the documents at hand while in the case of freely assigning phrases, the extracted phrases may or may not appear in a document. Since keyphrases concisely helps to summarize documents, they are used as a low-cost measure of detecting the similarities of documents (Frank *et al.*, 1999). Sarkar *et al.*, (2010) pointed out that keyphrases are used for several purposes which include text summarization, indexing and precision of search but not limited to these.

Keyphrase and keyword extractions are different in the fact that keyphrases that are extracted have one or more groups of terms while keywords have only one word. Keyphrases and keywords give a description of the documents to be analyzed but keyphrases give a better description of such documents. An extracted keyphrase list or keyword list associated with a particular document can act as a summary of the document to give a better description of such document that can assist readers or users of the documents to speed up their search of important information. Keyphrase extraction is an area in text mining that helps readers identifies vital groups of words in a document so as to conform to what they have in mind. Liu *et al.* (2009) stated that keyword extraction helps users to dive deeper into a document accurately in order to decide to read or not read such document.

In research articles, keywords help to give an indicator to the content of such document (Witten and Frank, 2000). Keyphrases helps as a description to a document because they reduce large text to meaningful phrases which can have different interpretation. The techniques used for selection range from the first occurrence of the word in document to how frequent the sentence is in relation to a document and how frequent such sentence is in other document (El-bethagy and Rafea, 2009). The parameters used in this technique are adjusted using the training set to have an accurate parameter. The only keyphrases that are formed are the one from the training phase; any possible phrase cannot be detected as far as it is not in the training set.

The extraction of keyphrases are carried out given a document at hand, the elementary lexical methods such as punctuations and stopwords removal help in extracting candidate phrases. The next operation is to check how frequent a particular phrase appears in a document. Another operation that can be considered is what part of the document a phrase appears i.e. in the title, abstract or in the body among other things.

This study is meant to improve on the drawbacks of graph-based keyphrase extraction by introducing the noun phrase identifier because most keyphrases are nouns or contain nouns (Awoyelu *et al.,* 2016). The work will also be using a different graph ranking algorithm to address the weak phrases that could be considered strong phrases and at the same time use a better clustering algorithm to solve the challenge of the k-means algorithm used. This study attempts to present a method that will automatically assign keyphrases to documents.

2.0 RELATED WORKS

Several approaches for unsupervised keyphrase extraction have been introduced in literatures. Some of these are based on statistics such as TFIDF while other approaches use complicated methods. Some approaches attempt to define what a keyphrase is by relying on certain statistical methods to understand the relationship between two candidate words of phrases. The more important a phrase is to a document, the more it is identified as quality keyphrase. The approaches to keyphrase extraction in documents are numerous. This is because several methods have been used for automating keyphrase extraction. These keyphrase extraction methods can be grouped into four which are: simple statistic approach, linguistic approach, machine learning approach and other approaches. They can be further collapsed into two main approaches (Beliga, 2014), which are supervised and unsupervised methods. The approaches to keyphrase extraction are depicted in Figure 1.

2.1 Supervised Approach

The supervised or machine learning approach involves a model being trained to identify a keyphrase. There are two groups of documents, one of the documents is used for training and the other is used for testing. This approach requires annotation in the training dataset making it stressful and not consistent. In most research articles, authors do' not assign keyphrases or keywords except they are asked to do it. A supervised approach uses training datasets, which consist of texts and the corresponding keyphrases that is assigned by human indexers or authors. The two main traditional methods that use the supervised approach are keyphrase extraction algorithm (KEA) (Frank *et al*., 1999) and Genitor Extraction also known as GenEx (Turney, 2000). KEA used a Naive Bayes classifier which was used to construct the term frequency and inverse document frequency (TFIDF) feature and the position feature of a phrase. GenEx used the genetic algorithm to construct a model that uses twelve (12) low-level parameters. The learning process adjusts the values of these parameters with each new entry from the training dataset to maximize the system's precision. The limitation of these techniques is that it was based strictly on machine learning i.e. if another dataset that is different from the training dataset is used, it misbehaves. Frank *et al.* (1999) further tested their Naive Bayes classifier with some other features. It was only the relative position and TFIDF that produced a better result. The features used in Frank *et al.* (1999) are used as the base features and have been applied similarly in other research (Hulth, 2003). KEA and GenEx had a good performance in extracting keywords but KEA was easier to adapt and has been the basic model for keyphrase extraction techniques.

Hulth (2003) used knowledge of linguistic like the syntactic analysis instead of just using the frequency of words as used in Frank *et al.* (1999). Yih *et al.* (2006) used a logistic regression classifier to identify keyphrases on webpages. This method used about ten attributes for calculating the features of a keyphrase. Nguyen and Kan (2007) used the TFIDF, position of the word and additional feature to extract keyphrases in scientific articles. The work used the position of a word or phrase i.e. if the word is found in the abstract or title among others, machine learning techniques from the neural networks to conditional random fields (CRF) have been proposed (Sarkar *et al.,* 2010; Zhang, 2008). Zhang (2008) used CRF to model features of a phrase. It made use of twenty features in understanding a keyphrase. A model for extracting keyphrases in meeting transcripts using the maximum entropy classifier was developed by Liu *et al.* (2011). Support vector machine (SVM) is also a machine learning model which can be useful for supervised machine learning. In SVM, a model is built to match the context of the testing documents. The supervised machine learning approach has been the most successful method of extracting keyphrases. This method of keyphrase extraction can be very demanding and time consuming on massive datasets (Awoyelu *et al.*, 2016).

2.2 Unsupervised Approach

Unsupervised approach does not use training dataset. Statistical features are used in some of these methods to check the importance of a term to a document. The two most important statistical features that have been used in different studies are the TFIDF and word frequency. Kireyev (2009) used the informativeness of a word to understand the semantics behind a particular word or phrase. Another important features used in unsupervised approaches are the clusteredness and inverse document frequency. The graph-based approach is another method used in unsupervised approaches where a document is modelled as graph where the words are represented as nodes and the relationship between those words are represented as edges. This modelled graph is now scored based on the technique used in calculating the node or edges. A very popular graph modelling method is the Textrank (Mihalcea and Tarau, 2004) which was adapted from google Pagerank (Brin and Page, 1999). Textrank is used to assign scores to element in a text document either the words or sentences in the document. Wan *et al.,* (2007) also proposed a graph based method that checks the relationship in a given document.

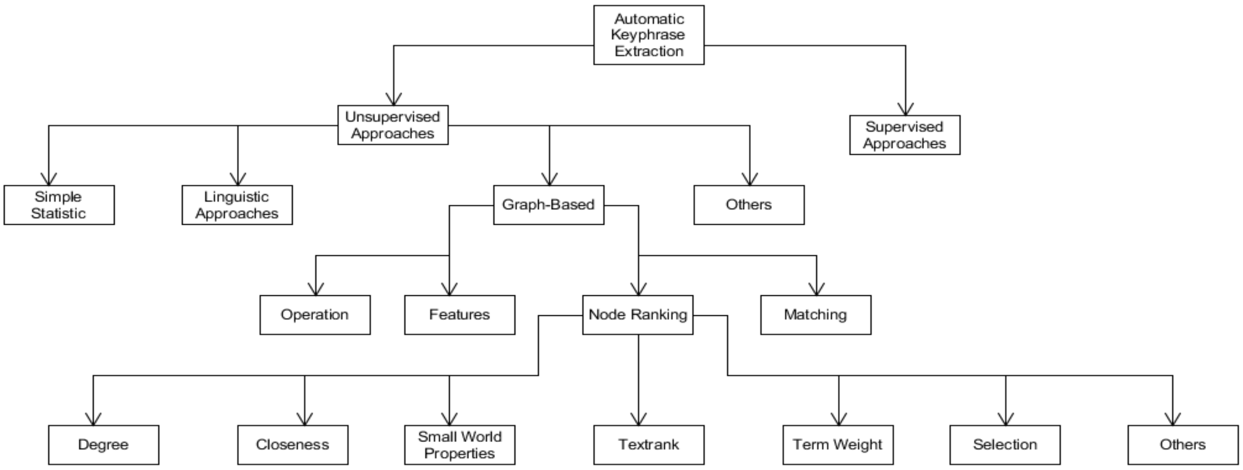


Figure 1: Keyphrase Extraction Methods (Beliga, 2014)

The relationship used in Wan *et al.* (2007) are words to words relationship, words to sentence relationship and sentence to sentence relationship. In response to the work of Wan *et al. (*2007), Wan and Xian (2008) developed Collabrank to improve the task of keyphrase extraction using the influence of several documents confined to a cluster. Unsupervised approaches do not use training dataset to extract keyphrases, they have a weakness in that they do not have an existing knowledge when compared with supervised approach. The weakness has been tackled using some hybrid approaches like HybridRank (Figueroa and Chen, 2014). HybridRank makes use of the textrank algorithm and KEA to produce a keyphrase. Awoyelu *et al.* (2016) also presented a system that uses the TFIDF, noun phrase identifier and a domain knowledge to tackle the training short comings of unsupervised approach. This model used the learning acquired from already extracted keyphrases to build a domain knowledge for future extraction.

Bhaskar (2013) presented a study on multi-document summaries employing automatic keyphrase extraction methods. The study proposed two approaches; the first approach deals with the automatic extraction of keyphrases while the second approach is concerned with the automatic generation of document summary using the keyphrases extracted. The machine learning method was used in this study which made use of CRF algorithm. A document graph-based topic/query focused automatic document summarizer was used for summarization where the extracted keyphrases were used as topic.

Figuerao *et al*., (2017) proposed an unsupervised approach called RankUp. The approach tried to enhance a graph-based keyphrase extraction approach by introducing an error-feedback mechanism that is synonymous to the idea of the backpropagation algorithm.

El bazzi *et al.*, (2017) used TextRank, which is one of the most successful graph-based approach in keyphrase extraction. It was used to calculate the weights of terms from documents (graphs). A node or vertice is used to represent the terms in the documents and the edges were used to represent term co-occurrences within a fixed window in the text. It was a novel unsupervised approach that was developed to extract keyphrases in Arabic document. The results of the study showed that the model when compared to KPMiner (El-bethagy and Rafea, 2009) performed better as both models are based in Arabic. The result showed that El bazzi *et al.*, *(*2017) outperformed Kpminer with a better precision and recall.

Biswas *et al.*, (2018) proposed an unsupervised graph-based approach for keyword extraction. The approach determines how vital a keyword is by measuring its influence on some parameters. The parameters in this study includes distance from central node, selectivity centrality, neighbouring nodes importance, node position and term frequency. Ideally, the study is based on node weight and node edge rank centrality. The model was validated using five different datasets which include America election dataset, harry porter dataset, IPL dataset, Uri attack dataset and Donald trump dataset. The result from the model was benchmarked with three existing models. The study observed that their proposed model outperformed others using precision and recall. The ranking approach used was not that efficient as it was proposed to use other ranking approach to enhance keyphrase extraction.

Yeom *et al.*, (2019) proposed a method that combines the statistical model, C-value method, and a graph-based model to improve the drawbacks of each model. The method for keyphrase extraction was developed by the graph-based model and the scores calculated by the new method are applied to the improved C-value method to estimate the final important scores of the keyphrase candidates. Subsequently, the model was evaluated using two datasets, SemEval 2010 and Inspec, and its results showed that the model outperformed the state-of-the-art models the existing graph-based ranking models.

Mallick *et al.*, (2019) proposed a graph-based text summarization model using modified textrank. The paper captured the aboutness of a text document. The method was developed using a modified textrank algorithm that is based on the concept of the pagerank defined for each page on the web pages. The method constructs a graph with sentences as the nodes and similarity between two sentences as the weight of the edge between them. Modified inverse sentence frequency-cosine similarity was used to give different weight to different words in a sentence whereas traditional cosine similarity treats the words equally. The graph method proposed outperformed the state of the art TF-IDF using precision, recall and f-measure as performance method.

2.3 Existing Graph-based Model

Yan *et al.,* (2017) proposed a graph-based method for keyphrase extraction. In the study, a document is divided into sentences and words. The obtained words are single characters other than phrases. Stop words are then removed from the words connection and all identical words are removed in the connection. The study did not tag a word to contain nouns as part of it because the other words without noun words have effect on sentence score. Each sentence and word represents a point in the graph. Three graphs were built which were the sentence to sentence graph, sentence to word graph and also the word to word graph. There exists an edge between every sentence or word which means every sentence or word is connected with each other. However, the sentences are not connected with all words due to the fact that every sentence has limited words and the words connection is composed of all the diverse words in the document. There are situations that a certain word does not appear in a sentence which means they cannot be connected The model is depicted in Figure 2. The Cosine similarity was used to identify the similarity between words and sentences in the word to word and the sentence to sentence graphs. The TFIDF was adapted to check the relationship in the sentence to word graph. The study used the single rank algorithm based on the work of Wan and Yian (2008), Yan *et al.,* (2017) added two assumptions to the singlerank. The first assumption was that a word must be important if it is connected to other important words and a sentence must be important if it is connected to other important sentences.

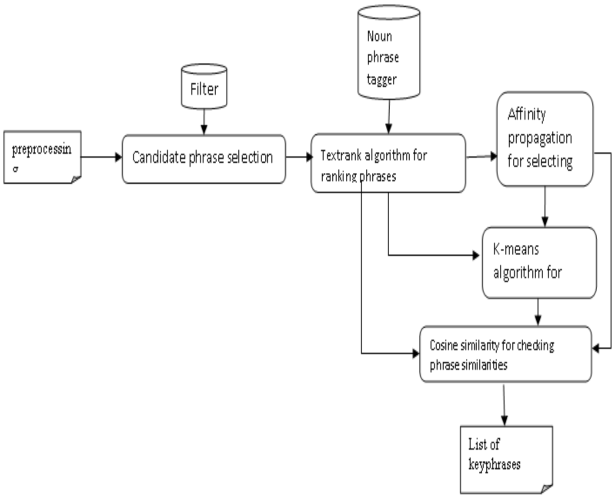

Noun phrase tagger
Filter
preprocessin σ
Candidate phrase selection
Textrank algorithm for ranking phrases
Affinity propagation for selecting
K-means algorithm for
Cosine similarity for checking phrase similarities
List of keyphrases

Figure 2: Keyphrase Extraction Model (Yan *et al.,* 2017)

The second assumption is a word must be important if it appears in many important sentences and a sentence must be important if it contains important words. The k-means algorithm was used to cluster the word to word graph whilst neglecting the other graphs as the aim was keyphrase extraction not text summarization. In evaluating the method in Yan *et al.,* (2017), The state of the art precision, recall and f-measure was used. The study made use of two dataset "Hulth 2003" and "500n dataset" to evaluate their method against TFIDF and textrank. The method has the best precision and f-measure score on the Hulth 2003 dataset and almost the same recall score with textrank. The method also has the best precision and recall value on the 500n dataset while textrank has the best recall value.

3.0 PROPOSED METHOD

The conceptual model of the proposed graph-based keyphrase extraction is depicted in Figure 3. In the pre-processing phase, the files to be used are converted to *.txt* files because the system accepts *.txt* files as inputs. The document is then filtered using list of stop words in English language like a, an, that. Punctuation marks such as comma and full stop are removed. The stopwords removed are replaced with spaces so as not to wrongly combine words, for example, processed information is technology, cannot be combined as “processed information” and “information technology” rather it should be combined “processed information” and “technology”. The “/” is used to replace full-stop sign and stop words so as not to combine the last word from a previous sentence with the first word of the next sentence.

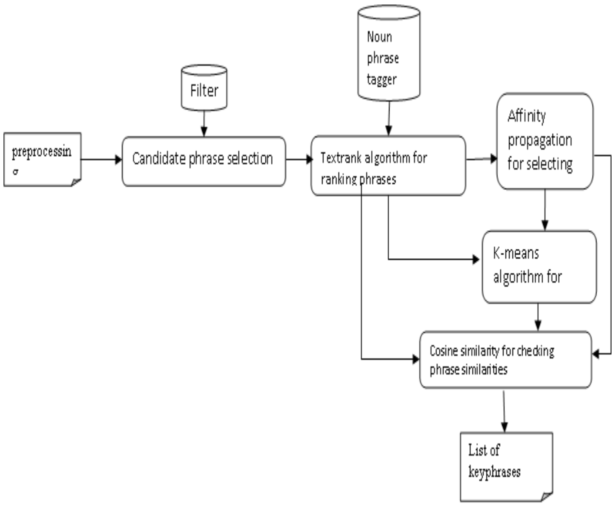

Noun phrase tagger
Filter
preprocessin σ
Candidate phrase selection
Textrank algorithm for ranking phrases
Affinity propagation for selecting
K-means algorithm for
Cosine similarity for checking phrase similarities
List of keyphrases

Figure 3: The Proposed Graph-based Model for Keyphrase Extraction

The phrase combination is limited to two words within a sentence as most author assigned keyphrases contain two words. After the words are combined, the phrases are built into the graph, the phrases are the nodes and the connection between phrases are the relationships i.e. phrases that are in a sentence are said to have a relationship while phrases in different sentences do not have any relationship. The textrank algorithm is applied to the graph built, every node has the same value initially and increases or decreases depending on the connection to it. After the textrank algorithm is applied, a score is attached to each phrase. The top n-rank are stored for comparison with the clustering algorithms. The graph built from the textrank algorithm is parsed into the affinity propagation algorithm for clustering and also to send its defined number of clusters to k-means algorithm. This is because affinity propagation selects its number of clusters automatically and k-means needs the user to select number of clusters. After k-means algorithm clusters the graph from the text rank algorithm.

Cosine similarity is used to check the closest cluster in k-means algorithm to the n-rank phrases from text rank and such cluster is selected as a quality cluster in k-means. Cosine similarity is also used to check the closest cluster in affinity propagation algorithm to the n-rank phrases from text rank and such cluster is selected as a quality cluster in affinity propagation. The intersection between the k-means cluster and affinity propagation cluster is selected as keyphrases. The block diagram for the conceptual model is as shown in Figure 4. After the text document is inputted into the system, the stopwords/punctuation are removed then the words are combined to a maximum of two words. The next phase is the graph building after which the phrase is tagged as noun or not then it is parsed into the textrank algorithkm. The output from the textrank is inputted into the affinity propagation, k-means and cosine similarity algorithm. Affinity propagation sends the clusters to the k-means to enable it select clusters. Affinity propagation sends the output to cosine similarity. K-means algorithm also sends its cluster to the cosine similarity. The cosine similarity checks the similarity between the two clusters and extracts a list of keyphrases. The conceptual model is an improved graph model for unsupervised keyphrase extraction where the noun phrase identifier and affinity propagation algorithm are combined to extract quality keyphrases. The combination of the noun phrase identifier and affinity propagation will help to improve the results of the keyphrases.

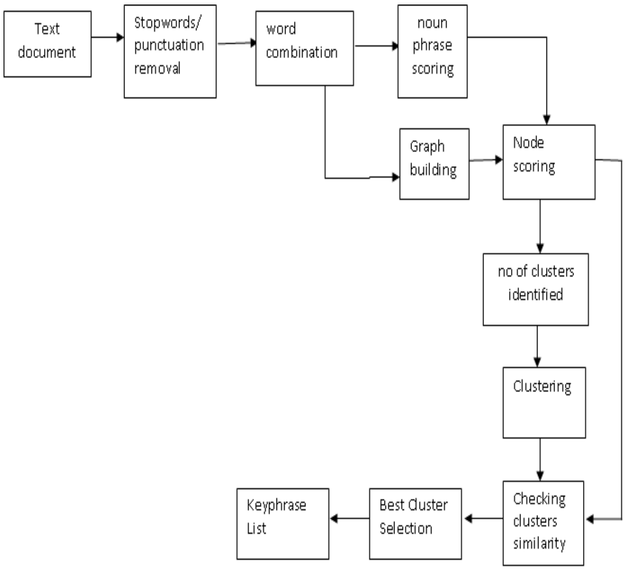


Figure 4: Block Diagram for the Proposed Graph-based Model

3.1 Techniques used in the Proposed Method

The proposed system is a utility that demonstrates an innovative and convenient way of extracting keyphrases without complexities. The system works in a way that allows it to extract keyphrase from documents independent of the domain of such document. The system accepts .txt as inputs and the words can also be copied to the interface. The resulting keyphrases will be fixed to ten for the system. The proposed keyphrase extraction uses some already known techniques such as stopword list, noun phrase tagger which are discussed as follows:

(1) Stopword list

Stop words are words which are filtered out before or after processing of natural language data (text). This contains a list of possible stop words and punctuations in English language. It covers a wide number of stopwords without getting too aggressive and including too many words which a user might search upon. The wordlist contains 429 words in English language. It contains articles such as “a”, “an” e.t.c. It also contains pronouns such as “he”, “she”, “they” e.t.c. It contains preposition such as “under”, “on”, “below” e.t.c. It contains conjunction and interjection such as “but “, “and” e.t.c. The stop words were extracted from (*http://www.lextek.com/manuals/onix/stopwords1.html*.

The stopwords extracted were added to the Natural Language Toolkit (NLTK) stopwords list so as to have a comprehensive list of stopwords.

(2) Noun phrase tagger

The noun phrase tagger tags a word or phrase to know whether such word is a noun or contain noun as part of it. A list of nouns is English language was extracted from the Oxford Advanced Learner’s Dictionary 7th Edition. The reason for the source of this nouns is because of the availability of the dictionary online. The noun tagger for this study uses noun list to check if a phrase contains a noun or is totally a noun. The dictionary is also freely and publicly available for download. The formula for calculating the noun phrase score is given in Equation 1

$$Ns = \frac{np}{wp} \qquad (1)$$

In calculating the noun phrase score, np is the number of nouns in a noun phrase, wp is the number of words in a noun phrase and Ns is the Noun phrase score.

(3) Textrank algorithm

After the words are assigned to contain nouns or not, so that only nouns phrases are considered. Then a graph of the document is created using phrase as a node and the relationship with the phrases as edges. The phrases (nodes/vertices) are denoted by V. Each phrase is connected to other phrases that are close to it in the text, this is represented by the connections on the graph denoted by E for edges. The algorithm is then run on the graph. Each node is given a weight of 1 then the algorithm goes through the list of nodes and collects the influence of each of its inbound connections. The influence is usually just the value of the connected vertex (initially 1, but it varies) and then summed up to determine the new score for the node, then these scores are normalized, the highest score becomes 1 and the rest are scaled from 0 to 1 based on that value. Each time through the algorithm gets closer to the actual "value" for each node, and it repeats until the values stop changing. This algorithm takes the top n-scored phrases that have been identified as important and outputs them for the cosine similarity as key/important phrases. The mathematical formula is given in Equation 2.

$$S(v_i) = (1 - d) + d * \sum_{j \in In(v_i)} \frac{1}{Out\ (V_j)} S(v_j) \qquad (2)$$

Given a vertex V_i, In(V_i) are set of vertices that point to it (predecessors) and Out(V_i) are set of vertices that V_i points to i.e. its sucessors. d is a damping factor that can be between 0 and 1 which was set to 0.85 (Brin and Page, 1998). The score for vi, S(vi), is initialized with a default value and is computed in an iterative manner until convergence.

(4) Affinity propagation algorithm

The graph of phrases built from textrank algorithm is sent to the affinity propagation algorithm to select the clusters. The algorithm is a clustering algorithm that exchanges messages between pairs of data points until a set of exemplars emerges, with each exemplar corresponding to a cluster. The Affinity Propagation algorithm takes as input a real number s(k,k) for each data point k—referred to as a “preference”. Data points with large values for s(k,k) are more likely to be exemplars.

There are two kinds of message exchanged between data points and each considers a different kind of competition. Messages can be combined at any stage to decide which points are exemplars and for every other point, which exemplar it belongs to. The “responsibility” r(i,k), sent from data point i to candidate exemplar point k, reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i, taking into account other potential exemplars for point i. The “availability” a(i,k), sent from candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar, taking into account the support from other points that point k should be an exemplar. r(i,k) and a(i,k) can be viewed as log-probability ratios. To begin with, the availabilities are initialized to zero: a(i,k) = 0. Then, the responsibilities are computed using the Equation 3.

$$r(i,k) \leftarrow s(i,k) - \max_{k' such\ that\ k' \neq k} \{a(i,k') + s(i,k')\} \qquad (3)$$

The responsibility update in Equation 3 let’s all candidate exemplars compete for ownership of a data point, the availability update, as shown in Equation 4, gathers evidence from data points as to whether each candidate exemplar would make a good exemplar.

$$a(i,k) \leftarrow \min\{0, r(k,k) + \textstyle\sum_{i' such\ that\ i' \in \{i,k\}} \max\{0, r(i',k)\}\} \qquad (4)$$

The availability a(i,k) is set to the self-responsibility r(k,k) plus the sum of the positive responsibilities candidate exemplar k received from other points. Only the positive portions of incoming responsibilities are added, because it is only necessary for a good exemplar to explain some data points well (positive responsibilities), regardless of how poorly it explains other data points (negative responsibilities). If the self-responsibility r(k,k) is negative (indicating that point k is currently better suited as belonging to another exemplar rather than being an exemplar itself), the availability of point k as an exemplar can be increased if some other points have positive responsibilities for point k being their exemplar. To limit the influence of strong incoming positive responsibilities, the total sum is thresholded so that it cannot go above zero. The “self-availability” a(k,k) is updated differently as shown in Equation 5.

$$a(k,k) \leftarrow \textstyle\sum_{i' such\ that\ i' \neq k} \max\{0, r(i',k)\} \qquad (5)$$

This message reflects accumulated evidence that point k is an exemplar, based on the positive responsibilities sent to candidate exemplar k from other points. The number of clusters is influenced by the preference values and the message-passing procedure. The n-rank phrases from the textrank algorithm is compared with each cluster in the affinity propagation. The cluster that contains most of the n-rank phrases is selected as the best cluster from affinity propagation. The criterion matrix for i and k is given in Equation 6

$$c(i,k) \leftarrow r(i,k) + a(i,k) \qquad (6)$$

(5) K-means algorithm

The clustering method known as k-means is a method that describes the ”best” partitioning of a data set containing k number of clusters. The method is defined by its objective function which aims to minimize the sum of all squared distances within a cluster, for all clusters. K-means algorithm is also known as the Lloyd’s algorithm. The objective function is defined Equation 7,

$$\arg\min E(C) = \textstyle\sum_{i=1}^{N} ||x_i - c_i\ \ ||^2 \qquad (7)$$

where C is the clusters ($c_{1,}c_{2,\dots}c_n$) and c_i is the centroid closest to the sample data point x_i. One aspect of k-means that makes it different from many other clustering methods is that the number of clusters is fixed when clustering occurs. This can be considered as both a weakness and a strength. One positive property of a fixed number of clusters is that the k-means method does not introduce new clusters in case of an anomaly data point, instead it sorts the anomaly data point to its closest cluster. The drawback of using a fixed number of clusters is that it might not be clear how many clusters a dataset might contain. Using an unsuitable k may cause the k-means method to produce poor results, possibly to the point of becoming unusable.

(6) Cosine similarity

The cosine similarity is a similarity measure that checks the degree of similarity of two entities, the entities could be words, sentences, documents etc. The cosine similarity used in this work calculates the similarity between the n-ranked phrases from the text rank algorithm with each clustering algorithm (Affinity propagation and k-means algorithm). The n-ranked phrases are assumed to be a cluster and is compared with each cluster of the two algorithms. The most similar clusters in affinity propagation and k-means algorithm are compare with each other to identify the similarity. The intersection of k-means algorithm and affinity propagation clusters are deemed keyphrases. The mathematical formula is given in Equation 8,

$$cosine\ similarity = \frac{x_i\, x_j}{||x_i||\,||x_j||} \tag{8}$$

where xi and xj are the two entities, the cosine similarity takes the product of the two entities divided by the product of the magnitude of the entities to measure the similarity.

3.2 Proposed Model Formulation

Existing 'techniques to be used in formulating the proposed model includes the textrank algorithm, noun phrase identifier, affinity propagation, k-means algorithm and cosine similarity. There are two approaches to the proposed model. The first approach is the Ranking approach and the second approach is the clustering approach. The ranking approach is formulated using the textrank algorithm and noun phrase identifier while the clustering approach is formulated using the k-means algorithm and affinity propagation. The cosine similarity is only used to compare the ranked phrases and the clustered phrases. The formulation approaches are discussed in the following section:

(1) Ranking approach

The ranking approach uses the textrank algorithm and noun phrase identifier in formulating the ranking model. After the words have been combined to a maximum of two words. The graph of the phrases connection is plotted. The score of an individual phrase is calculated using the noun phrase identifier given in Equation 1 and textrank algorithm in Equation 2. The new score of each phrase is computed as shown in Equation 9.

$$Phrase\ Rank\ Score = \ S(v) + Ns \tag{9}$$

where S(v) is the phrase score by textrank after convergence and Ns is the noun phrase score and Phrase rank score is the score of each phrases after computation. Each phrase is ranked in descending order of phrase rank score. The top n-rank phrases are assumed to be a cluster C_{tn}.

(2) Clustering approach

The clustered phrases from the ranking approach is stored and compared with the two clustering algorithm. The affinity propagation algorithm takes the graph plotted by textrank and clusters it based on its algorithm, the resulting clusters is sent to k-means to enhance its clustering. From Equation 5 the number of cluster is equal to the number exemplars. The number of cluster from the affinity propagation is represented with nC_a which is number of cluster gotten from affinity propagation. The cluster groups in affinity propagation is defined as shown in below.

$$C_a \leftarrow c_1, c_{2,\dots\dots,} c_n \tag{10}$$

where c1, c2 and cn are the individual clusters in affinity propagation. The similarity between affinity propagation and the ranking approach is calculated using cosine similarity where C_{tn} is the phrases from the ranking approach and C_a is the clusters from affinity propagation. Using Equation 7 to calculate the cosine similarity between each clusters which is given in Equation 11.

$$\text{Ctna} = \arg max \frac{C_{tn}\, C_a}{||C_{tn}||\,||C_a||} \tag{11}$$

The cluster with the highest cosine similarity is deemed the new cluster Ctna. The k-means takes as the input the graph from textrank and takes the number of clusters from affinity propagation nC_a. it uses the cosine similarity from Equation 7 to compare its cluster. The k-means cluster is represented with $C_{k.}$ The maximum cosine similarity between k-means cluster and textrank cluster is given in Equaton 12.

$$Ctnk = \arg max \frac{C_{tn}\, C_k}{||C_{tn}||\,||C_k||} \tag{12}$$

The cluster with highest cosine similarity is deemed the new cluster Ctnk. Finally, the intersection between the k-means cluster and affinity propagation is regarded as the keyphrases which is given in Equation 13.

$$keyphraselist = C_{tna} \cap C_{tnk} \tag{13}$$

4.0 RESULTS

The design of the proposed model was carried out using the Unified Modeling Language (UML). UML is a standard visual modeling language used for modeling business processes and other similar processes. It is used for analyzing, designing so as to implement software based systems (Bell, 2003). UML diagrams include use cases, activity diagrams, class diagrams e.t.c. This study focused on the use case and the activity diagram. They are discussed below:

(1) Use case diagram

Use case diagrams are used during requirements elicitation and analysis as a graphical means of representing the functional requirements of the system.

Proposed Algorithm

Pseudocode for the Proposed Method

```
Input: Document D = {w1,w2,...,wn}, S = {stopwords_list}, Q = {punctuation}, N = {NounList}
Replace S and Q from D with "/"
Combine w in D until "/"
D = {p1,p2,...,pn}
    If N in D then
        NS = no of nouns in phrase / no of words in phrase
    end if
G ?  BuildGraph(D)
Scores ?  (1.0,1.0,....1.0) // initialize scores
converged ?  false
    while converged == false do
    converged ?  true
    old scores ?  scores
        for phrase p ? 1,2,...,length(D): do
        // update phrase score according to rule given above
        Scores [Phrases ?  updatePhrase(G, Phrase, d=0.85, scores)
            if |scores[phrase] - oldscores [phrase]| > ε then
                converged ?  false
            end if
        end for
    end while return TS // phrases with textrank scores
            TN = TS + NS // textrank scores and noun phrase scores.
given two datapoints i and k // these are the TS scores of different phrases
    result1 = -?
    for each data point z such that (z ? k):// z is actually not i or k
        temp1 = avail[i,z] + simi[i,z]
        if(temp1 > result);
            result1 = temp1
    resp[i,k] = simi[i,k] - result1 // responsibility matrix
if (i?k):
    sum = 0
    for each datapoint z such that (z?i) and (z?k):
        temp2 = resp[z,k]
        if(temp2 > 0):
            sum = sum +temp2
    result2 = sum + resp[k,k]
    if (result2 > 0):
        result2 = 0
else:
    sum=0
    for each data point z such that (z?k):
        result2 = resp[z,k]
        if(result2 > 0):
            sum = sum + result2
    avail[i,k] = sum // availability matrix
crite[i,k] = resp[i,k] + avail[i,k] /* criterion matrix: objects having similar results in their row are in the same exemplars*/
AP = ($AP_1$, $AP_2$,...,$AP_n$) // no of exemplars or clusters
Arbitrarily choose AP objects from TS as the initial clusters centers;
Assign initial values for $AP_1$, $AP_2$, ..., $AP_n$;
Repeat:
1.    (re)assign each object to the cluster to which the object is most similar, based on the    mean    value of the objects in the clusters;
2.    Update the cluster means, i.e., calculate the mean value of the objects for each cluster
Until no change in clusters
KM = $KM_1$,$KM_2$,...$KM_k$ //Kmeans cluster
Maximum cosine similarity1 = (TN * AP) / ( ||TN||*||AP||
Maximum cosine similarity2 = (TN * AP) / ( ||TN||*||AP||
```

Use cases are developed during requirements elicitation and are further refined and corrected as they are reviewed during analysis. The use case diagram for this study is depicted in Figure 5, an actor is shown as the agent who interacts with the system. This use case diagram shows the keyphrase extractor consisting largely of two users. The user provides the input to the system and can interrupt an extraction if he/she decides to change or make any adjustment to the input of the system. The user can also see the list of keyphrases extracted. The keyphrase extractor extracts keyphrases, accepts adjustment made on the document input, prints out the list of keyphrases extracted and also collects the text input from the user for analysis.

(2) Activity diagram

Activity diagrams are used to describe the dynamic aspects of a system's design. They are basically flowcharts representing the flow from one activity to another. They described the operations of the system. Activity diagrams are used during the design phase of complex methods. Alternately, the activity diagram can also be used during analysis to break down the complex flow of a use case. An activity diagram has been used to describe the various activities that in the proposed model as shown in Figure 6. The text document is inputted into the system, stopwords are removed, the words are combined to a maximum of two words, the graph of the phrase connection is built before the textrank algorithm is applied to it, the phrase is checked if it contains a noun or not, the score of the phrase is ranked. The graph built from textrank is sent to affinity propagation for clustering then the number of clusters detected by affinity propagation is sent to k-means algorithm. Cosine similarity finally makes a comparison between the textrank and the two clustering algorithm to select the cluster with the maximum similarity.

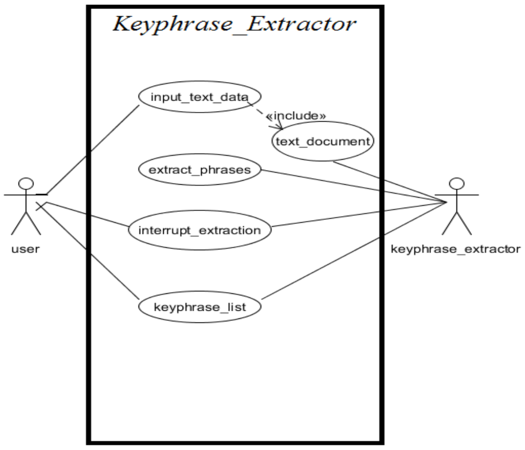


Figure 5: Use Case Diagram for the Proposed Model

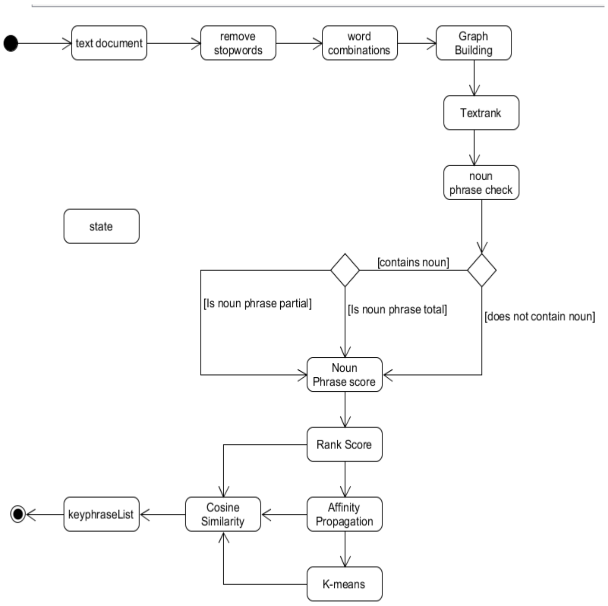


Figure 6: Activity Diagram for the Proposed Model

5.0 CONCLUSION

Keyphrases are phrases that represent a document at hand. The study proposed a conceptual graph model for extracting keyphrases in document. The graph model represents the document as a graph and as such a phrase as a node and the relationship with such nodes as edges. Existing methods that have used the graph-based model did not take into consideration of noun phrases. The k-means algorithm used as a method of clustering is a very good clustering algorithm provided the number of clusters is known in advance but this has a setback on keyphrase extraction task as the number of clusters is unknown before extraction. The affinity propagation proposed is a clustering algorithm that determines clusters based on their responsibilities and availabilities to each other. This algorithm helps to solve the challenge with k-means by identifying the possible number of clusters. The study is limited to document extraction in articles and proposes a model with better precision, recall and f-measure scores when compared with the state of the art models. The conceptual graph-based model could further be developed so as to compare with existing models for keyphrase extraction using precision, recall and f-measure as performance metrics.

REFERENCES

Awoyelu I. O., Abimbola R.O., Olaniran A.T., Amoo A.O and Mabude C.N. (2016). Performance Evaluation of an Improved Model for Keyphrase Extraction in Documents. *Computer Science and Information Technology* 4(1):33 - 43. Doi: 0.13189/csit.2016.040106.

Beliga, S. (2014). Keyword extraction: a review of methods and approaches. *University of Rijeka, Department of Informatics*, 1-9.

Bell, D. (2003). An Introduction to the Unified Modeling Language. *IBM Global Services. Available at: http;//www.therationaledge.com/content/jun_03/f_umlintro_db.jsp.* Accessed: 20th December 2018.

Biswas S., K., Bordoloi, M. and Shreya, J. (2018). A Graph-based keyword extraction model using collective node weight. *Elsevier Expert Systems with Applications*, pp 51-60. https://doi.org/10.1016/j.eswa.2017.12.025.

Bhaskar, P. (2013). Multi-document Summarization Using Automatic Keyphrase Extraction. *In Proceedings of the Student Research Workshop associated with RANLP 2013*, Hissar, Bulgaria. pp. 22-29.

Bozlu, B., and Demirörs, O. (2008). A Conceptual Modeling Methodology: from Conceptual Model to Design. *In Proceedings of the 2008 Summer Computer Simulation Conference.* Society for Modeling & Simulation International. p. 30.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 107-117.

El-Bazzi, M., S., Mammass, D., Zaki, T. and Ennaji, A. (2017). A graph-based ranking model for automatic keyphrases extraction from arabic documents. *Springer International publishing AG 2017. pp. 313-322.* DOI:10.1007/978-3-319-62701-4_25

El-bethagy, S., R. and Rafea, A. (2009). Kp-miner: A Keyphrase Extraction System for English and Arabic Documents. *Elsevier Information systems. 34(1), 132–144.*

Frank, E., Paynter, G. W. andWitten, I. H. (1999). Domain-specific keyphrase extraction. *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (IJCAI). pp 668-673

Figueroa, G. and Chen, Y.S. (2014). Collaborative ranking between supervised and unsupervised approaches for keyphrase extraction. *ROCLING '14: The 2014 Conference on Computational Linguistics and Speech Processing*.

Figueroaa, G., Chena, P. and Chena Y. (2017). RankUp: Enhancing Graph-Based Keyphrase Extraction Methods with Error-feedback Propagation. *Computer Speech & Language.* 47, 112-131. doi: 10.1016/j.csl.2017.07.004.

Hulth, A. (2003) Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *In proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003),* 216-223.

Kireyev, K. (2009). Semantic-based estimation of term informativeness. In: Proceedings of Human Language Technologies: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics*, pp. 530–538.

Liu, F., Pennell, D., Liu, F. and Liu, Y. (2009). Unsupervised Approach for Automatic Keyword Extraction using Meeting Transcripts. *ACL*. pp. 620-628.

Liu, F., Liu, F., and Liu, Y. (2011). A supervised framework for keyword extraction from meeting transcripts. *IEEE Trans. Audio, Speech & Language Processing*, 19(3):538–548.

Mallick, C., Das, A., K., Dutta, M., Das, K., A. and Sarkar, A. (2019). Graph-based Text Summarization Using Modified TextRank. *Soft Computing in Data Analytics, Springer Nature,* PP 137-147.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. *In Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP)*, pp. 404-411.

Nguyen, T. and Kan, M. (2007). Keyphrase extraction in scientific publications. *Proceedings of ICADL2007.*

Sarkar, K., Nasipuri, M. and Ghose, S. (2010). A new approach to keyphrase extraction using neural networks. *International Journal of Computer Science Issues. Vol 7, pp. 16-25.*

Shi, W., Zheng, W., Yu, J., X., Cheng, H. and Zou, L. (2017).Keyphrase extaction using knowledge graphs. *Springer international publishing AG 2017, APWeb-WAIM 2017, part 1, LNCS*

10366, pp. 132-148, 2017. DOI: 10.1007/978-3-319-63579-8 11.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information retrieval 2 (4),* pp 303–336.

Witten, I., H. and Frank, E. (2000) Data mining: Practical machine learning tools and techniques with Java implementations. *Morgan Kaufmann, San Francisco, CA*.

Wan, X., Yang, J. and Xiao, J. (2007) Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic*, pp 552–559.

Wan, X. and Jianguo, X. (2008). Single document keyphrase extraction using neighborhood knowledge. *In Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp 855–860.

Yan, Y., Tan, Q., Xie, Q., Zeng, P. and Li, P. (2017). A Graph-Based Approach of Automatic Keyphrase Extraction. *Procedia Computer Science 107.* pp. 248-255.

Yeom, H., Ko, Y., and Seo, J. (2019). Unsupervised-learning-based keyphrase extraction from a single document by the effective combination of the graph-based model and the modified C-value method. *Computer Speech & Language*, *58*, 304-318.

Yih, W., T., Goodman, J. and Carvalho, V., R. (2006). Finding advertising keywords on web pages. *In: Proceedings of the 15th international conference on World Wide Web. ACM*, pp. 213–222.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems.* 4(3), 1169-11.

CFD ANALYSIS OF THE THERMAL ENHANCEMENT OF PEM FUEL CELL COOLING CHANNEL USING DIFFERENT CROSS SECTIONAL GEOMETRY

***Obayopo S. O., Atofarati E. O., Oyekunle M. A., Akinpelu A. O. and Oyelami O.**
Department of Mechanical Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria,

*Email of Corresponding Author: obayopos@oauife.edu.ng/sirajolanre@yahoo.com

ABSTRACT

This paper aims at numerically investigating the performance of a 3D model of the PEM fuel cell using different cross-sectional geometry for a serpentine flow channels. PEM fuel cell have been researched on and observed to be a promising solution to clean energy. Thermal optimization at operating condition for efficient power production has been a major difficulty encountered in their design. In this paper, a three dimensional, single phase, constant heat flux and transient state model of a typical Proton Exchange Membrane (PEM) fuel cell cooling channel was investigated numerically. The work mimics a typical PEM fuel cell having its cooling plate between the bi-polar plates. The continuity equations, momentum equations and energy equations were solved using a Computational Fluid Dynamics (CFD) tool called ANSYS Fluent having embedded FEM solver. Circular, triangular and squared cross-sections were considered in this work with an aim of identifying the best design for typical PEM fuel cell for optimum performance. The result of this study revealed that the squared cross section is the most effective for heat dissipation in the system. Hence, the efficiency of the fuel cell could be improved by using squared cross sectioned cooling channel rather than the conventional circular.

Keywords: PEM fuel cell, computational fluid dynamics, cooling channels, bi-polar plates, optimum performance.

NOMENCLATURE

C_P	[J/kg. K]	Specific heat capacity of coolant at constant pressure
k	[W/mK]	Thermal conductivity
q	[W/m^2]	Heat generation per unit surface area
f	[N]	Force
T	[K]	Temperature
x	[m]	Cartesian axis direction
y	[m]	Cartesian axis direction
z	[m]	Cartesian axis direction

Special characters

ρ	[Kg/m^3]	Density of the coolant
μ	[m^{2}s^{-1}]	Dynamic viscosity of coolant
τ	[Pa]	Shear Stress

1.0 INTRODUCTION

Energy generation is a key factor for Industrialization and a major stimulant for good standard of living in any society. However, power generating plants have been observed to significantly contribute to air pollution, as it's the foremost source of greenhouse gases. Hence, the quest for clean energy has become the state-of-the-art in the world right now. Research into cheap, clean and renewable energies; solar energy, wind energy, fuel cells among others has recently received tremendous attention owing to this growing concerns for continual consumption, limited nature and adverse impact of fossil resources, which constitute the world's main energy source, on the environment [1].

Among the different categories of fuel cells, Proton Exchange Membrane (PEM) fuel cell as in Fig.1, have the overriding advantages of rapid start up, low operating temperature and high power density. Hence, PEM fuel cells are used predominantly for transportation, stationary and portable power generation [1], [2]. In PEM fuel cells, electrochemical reaction takes place and electrical energy is obtained directly from the chemical energy in the fuel without actual combustion or fume. The electrical energy released is as a result of the hydrogen oxidation reaction and oxygen reduction reaction at the anode and cathode catalyst layers, respectively [2].

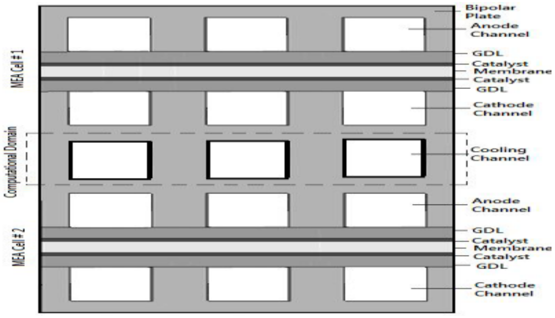


Fig 1. Complete stack of PEM fuel cell. [3]

2.0 LITERATURE REVIEW

Several advances have been made to improve the efficiency of PEM fuel cells, however durability and cost still prove as the major factors limiting its commercialization [2], [4]. A number of numerical studies have been carried out to improve the technology behind PEM fuel cell [4], [5]. The performance, stability, and durability of PEM fuel cells largely depend on the water and thermal management of the fuel cells [6], [7] .

While in operation, the membrane of the fuel cell must be well humidified, to reduce the proton transport resistance associated with water content and local temperature. However, excessive water can obstruct the transport passages in the permeable regions and thus give rise to water flooding problem. Moreover, a significant amount of heat is generated continuously by electrochemical reactions. within PEM fuel cells, and it is in the same order of magnitude as the electrical power generated, that is, the heat loss in PEM fuel cells is the same in magnitude or somewhat higher than the electrical power generated. To reduce the maximum temperature and avoid overheating of fuel cells, the heat produced must be effectively exhausted using appropriate cooling channels and methods. A lower temperature may result in water flooding while higher temperature increases the electrochemical reaction rate and may also lead to membrane dehumidification and performance degradation, hence it is compulsory to cautiously control the temperature and maintain a small temperature difference in a PEM fuel cell. In order to meet the requirement of heat dissipation from fuel cells stack system, an effective cooling system is necessary[8].

In large-scale PEM fuel cell stacks, liquid cooling technique is usually employed as a means of cooling because of its boundless cooling capacity compared to air. The liquid water is circulated through the flow channels of cooling plates (cooling channel) in the fuel cell stacks[3], [9]–[19]. The cooling plates serve as uniform temperature distributor and absorber in PEM fuel cells by reducing maximum temperature so as to ensure homogeneous electrochemical reaction in the fuel cell stacks. Thus, enhancing the durability, operational stability and performance of the PEM fuel cells. More uniform temperature distribution in fuel cell stacks can easily be achieved by increasing the flow rate of the coolant circulating through a cooling plate. On the other hand, more pumping power is needed for higher flow rate and this can, as a result adversely affect the overall efficiencies of PEM fuel cells owing to parasitic power consumption. As a substitute method, a number of coolant flow field designs and channel reforms have been recommended and tested for improved cooling performance of PEM fuel cells [3], [11], [12], [15], [18], [19].

In line with these studies, Chen *et al* [13] gave account of the performance of cooling plates with various serpentine and parallel flow fields. It was noted that the pressure drop of parallel channel design is lower than that of serpentine flow design. The results also revealed that the cooling effect of the serpentine flow channel is more enhanced than that of parallel channel design. The performance of cooling channels with several multi-pass serpentine flow fields adopted as coolant channels was investigated by Yu *et al*. [13]. In the study, the effects of heat flux and flow velocity on the cooling effect were also investigated and compared. It was reported that the cooling effect of the multi-pass serpentine flow fields is better compared to the conventional serpentine flow field. A similar study was also carried out by Baek *et al*. [16]on the performance of cooling channels with parallel, serpentine and multi-pass serpentine flow fields and their results revealed that there would be a considerable improvement in the uniformity of temperature in multi-pass serpentine flow fields if the coolant flow rate of parallel paths is properly distributed.

Kurnia et al [20] conducted analytical study on the heat transfer performance of eight cooling channel designs including parallel, serpentine, wavy, oblique fin, and coiled-based channels and the advantages and disadvantages of each of the designs were discussed in detail. The cooling of PEM fuel cell stacks, with air as the cooling medium, was conducted using the divided serpentine and spiral flow channel designs recommended by Ravishankar and Prakash [15]. The results obtained showed that the thermal performance was more enhanced by employing the new designs at all Reynold numbers.

Afshari et al. [17] examined the performance of cooling plates with parallel, serpentine and metal foam porous media flow field designs. Numerical and experimental investigation of the thermal performance of chaotic geometries proposed by Lasbet et al. [18], [19]. It was established that there is a significant improvement in the convective heat transfer when the chaotic geometries are adopted. The fluid flow and heat transfer characteristics of a novel zigzag-shaped flow field used as the cooling plates were numerically investigated and compared by Afshari et al. [3] to the conventional straight channel flow field design.

The results obtained led to the conclusion that the cooling performance of fuel cells can be improved by employing the zigzag channel flow field design as the coolant fluid distributors. In a recent study, Li et al. [11] examined and analyzed cooling plates with novel non-uniform flow channel designs and conventional straight channel flow field design. The influence of heat flux and inlet Reynolds number on the cooling performance were also examined. It was established that there is a significant improve-ment in the cooling performance as the recommended flow channel designs are used.

This study numerically investigated the perform-ance of a single-pass serpentine flow fields as coolant flow field designs section of a large-scale cooling plate with 22mm x 22mm square plate. A commercial CFD code, ANSYS Fluent, was used to simulate the detailed heat transfer and fluid flow in the cooling plates using different coolant flow field designs. Three different geometries of the conventional serpentine flow fields considered for comparison purposes are: circular, triangular and square. The cooling performance of each coolant flow field design was quantitatively assessed in terms of the maximum temperature, temperature uniformity, and coolant pressure drop characteristics. Furthermore, the effects of inlet Reynolds and heat flux and number on thermal performance are investigated

3.0 MATHEMATICAL MODEL

The three dimensional PEM fuel cell model used in this work was solved as a lamina, transient state and single phase problem. The model is a mimic of an ideal single cooling plate (bipolar plate) in a full stack of PEM fuel cell as in figure 1. The computational domain of the fuel cell stack considered under this study is presented in figures 2(a, b, c), the bipolar plates been the top and bottom surfaces of the geometrical model. The geometrical model consist of the cooling plate, adiabatic wall, symmetry wall, coolant inlet, coolant outlet as shown in figure 2. The bipolar plate is considered to be a heat generating surface with uniform heat flux (q) of 100W/m^2 and the heat is removed by conduction through it, then to the coolant. An examination of the effect of triangular, circular and rectangular serpentine cooling channels on a section of the PEM fuel cell with a domain dimension of 22mm × 22mm × 3mm was carried out. The coolant in this work is liquid water while the bipolar plate is made of copper. The surrounding temperature and pressure is taken to be 300K and 1bar respectively. Characteristic length of 1mm was used for the triangular and squared cross sections which corresponds to hydraulic diameter of 1mm for the circular cross sectional geometry.

All the simulations in this work were performed on the basis of the same inlet velocity, material properties and parameters as described in Table 1.

3.1 Governing Equation

The mathematical equations utilized and solved by the CFD tool in an embedded form in this work include, he Continuity equation, Momentum equation, Energy equation, Fourier law of conduction and the Newton's law of cooling among others, as in [21].

Continuity Equation

The continuity equation is given in the conservative form in equation 1.

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho V) = 0 \qquad (1)$$

Momentum Equation

The conservative form of the momentum equation in the x- component, y- component and z- component solved
by the CFD tool is presented respectively in equation 2.

$$\frac{\partial(\rho u)}{\partial t} + \nabla.(\rho u V) = -\frac{\partial p}{\partial x} + \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z}\right) + \rho f_x$$
$$\frac{\partial(\rho v)}{\partial t} + \nabla.(\rho u V) = -\frac{\partial p}{\partial y} + \left(\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{zy}}{\partial z}\right) + \rho f_y$$
$$\frac{\partial(\rho w)}{\partial t} + \nabla.(\rho u V) = -\frac{\partial p}{\partial z} + \left(\frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z}\right) + \rho f_z \qquad (2)$$

Energy Equation

The energy equation is given by;

$$\rho Cp \frac{DT}{Dt} = \nabla.k\nabla T + \frac{Dp}{Dt} + \mu\Phi \qquad (3)$$

Fourier Law of Conduction

The Fourier law expression is given in the reduced form as:

$$Q_{Cond} = -kA\frac{\Delta T}{\Delta x} \qquad (4)$$

Newton's law of cooling

$$Q = hA(T_s - T_\infty) \qquad (5)$$

Reynolds number

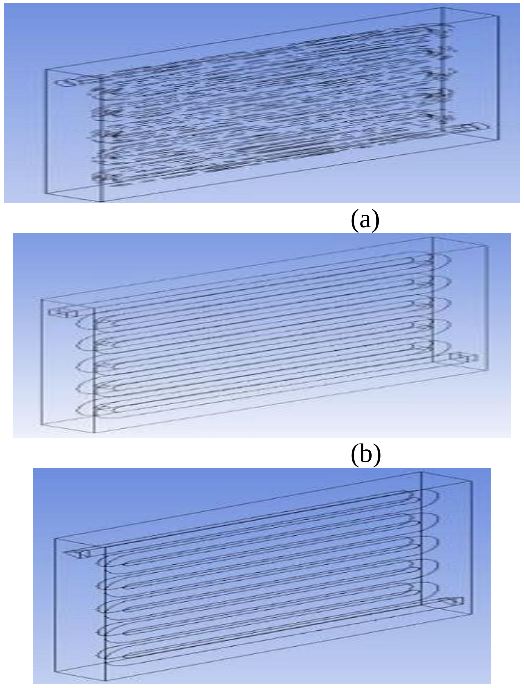
For incompressible fluid Reynolds number is given as

$$Re = \frac{\rho u D}{\mu} \qquad (6)$$

Table 1: Cooling plate parameters for PEM fuel cell

Parameter	Value
Heat generation per fuel cell (W)	100
Active area (m^2)	0.000484
Length of Plate (mm)	22
Width of Plate (mm)	22
Plate Thickness (mm)	3
Characteristics Length (mm)	1
Cooling fluid	Water
Inlet velocity (ms^{-1})	1.49 – 7.5
Inlet temperature (k)	300
Solid plate material	Copper

The Cartesian coordinate position of the fluid is described as x, y & z; while its flow velocity is described as u, v & w defines the in the x, y & z direction respectively; T is the temperature with respect to position of the fluid; ρ is the density of coolant {i.e. water} (998.2 kg/m^3); coolant was assumed to be incompressible, and μ is the viscosity of water 300K (1.003e-3).


(a)

(b)

(c)

Figure 1. The 3-D PEM Fuel Cell Geometrical model (a) Circular (b) Squared & (c) Triangular cross section

3.2 Numerical Method

PEM fuel cell with control volume of 1.4508e-6 m³ was used in this work. The geometry was modelled using SolidWorks, saved as STEP files and imported into the ANSYS design modeller interface. The geometry was discretized and solved, taking into account the mesh independence study. Suitable discretization based on accuracy and low cost of solution was used with 415502 element and 1100205 nodes. The mesh was then imported into the ANSYS fluent interface where the governing equations described for this heat transfer problem was solved extensively. The pressure-based solver was used with energy model put on. Water (liquid state) and copper (solid state) were selected for the materials.

3.3 Boundary Conditions

Boundary conditions were set thus: The Inlet was set as velocity-inlet with a velocity magnitude being a function of the Reynolds number. (1.46m/s corresponding to Re=100, 2.19m/s corresponding to Re=150, 2.92m/s corresponding to Re=200, 3.65m/s corresponding to Re=250, 4.38m/s corresponding to Re=300, 5.84m/s corresponding to Re=400, 7.50m/s corresponding to Re=500,). The inlet temperature of the coolant (water) is 300K. The Outlet was set as a pressure-outlet with zero gauge pressure. Backflow total temperature was set as 300K; while the actual outlet temperature in this study is a function of the cross section geometry and the Reynolds number of the coolant.

The Bi-polar plate (top and base of the PEM fuel cell) is a copper plates walls with a heat transfer rate Q = 100 W. The plates are the source of heat as they have direct contact with the anode and cathode in a full PEM fuel cell. The front and back walls were set as adiabatic wall; with no heat transfer across them base on the nature of lamina flow the no-slip condition was used at the walls. The side walls were set as symmetries; indicating the extensibility of the PEM fuel cell.

Convergence condition in this solution was set by default such that the sum of relative residuals was less than 1.0e-3. The thermal flow calculation was done within a flow time of 1 seconds; using a time step size of 0.001s and 100 times step. The cooling channels used in this study were both serpentine and modified serpentine channels as shown in the Fig 3 and 6. All the numerical calculations and solution for this work were done using ANSYS fluent on laptop with AMD E-300 1.32GHz.

4.0 RESULT AND DISCUSSION

To improve the performance of PEM fuel cells, efficient cooling systems were used by optimizing cross sectional area and fluid flow design. A modified serpentine design was also investigated. This design had been carefully selected to ensure uniform heat distribution on the plate, minimizing hot spots.

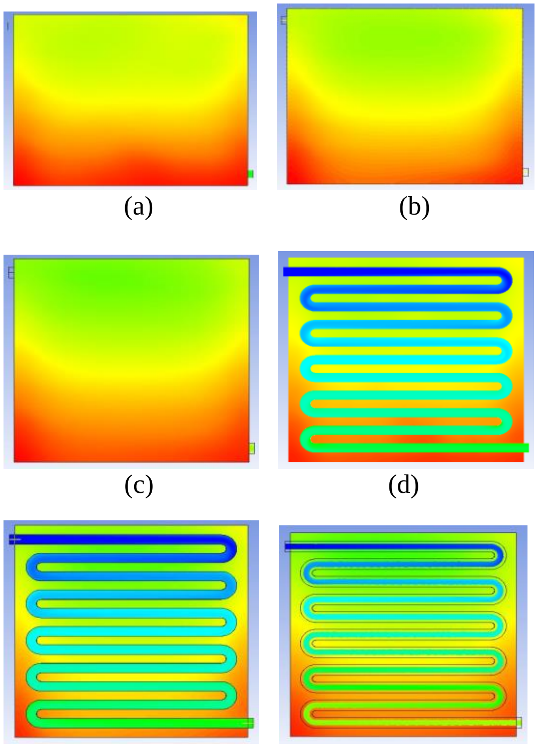
Figure 3 shows the temperature contours for the cooling plate for the serpentine cooling channel profiles at X-Y mid-plane. The temperature variation was observed from the top ‘coolant inlet’, to the base ‘coolant outlet’. The temperature of the coolant at the inlet is usually low and gradually increase as it flows through the cooling channel. This however this leads to the formation of ‘hot spots’ at the lower ends of the cooling plate.

It was observed from the temperature contour in Figure 3 and the graph of temperature to Reynolds number for respective cross sections in Figure 4; that the square cross sectioned plate has the lowest maximum temperature among the plates. The square cross section performed best, then the circular followed by the triangular flow channels. This performance was attributed to higher cross sectional area of the square cross section for the fixed characteristics length.

Generally, the maximum temperature decreases as the Reynolds number increases. The modified serpentine design as seen in Fig 6 tends to have a better temperature distribution and less concentrated hot spot when compared to the conventional serpentine design.

In the modified serpentine design, temperature variation of the square cross section had the least red spot as seen on the graph in Fig 7. Also, it takes after the conventional serpentine that has its square cross section as the best performing cooling plate. This performance which helped to validate that flow channel with square cross sectional area performs better. It was also observed that the new serpentine design has an over all difference better than the conventional design (though little).

However, increased pressure drop across the modified serpentine flow channel was a major concern. In Figure 5 the circular cross section was observed to have the least pressure drop in when compared to the square and triangular cross section. A higher pressure drop neccessitate for more pump power for transportation of the cooling fluid.


(a) (b)

(c) (d)

(f)

Figure 3: Temperature contour along the top surface and temperature contour at the mid plane.

Figure (a), (b) & (c) are the temperature profile of the circular, square and triangular cross section respectively; (d), (e) & (f) are the temperature contour at the mid plane of the circular, square and triangular cross section respectively at Reynolds number of 500

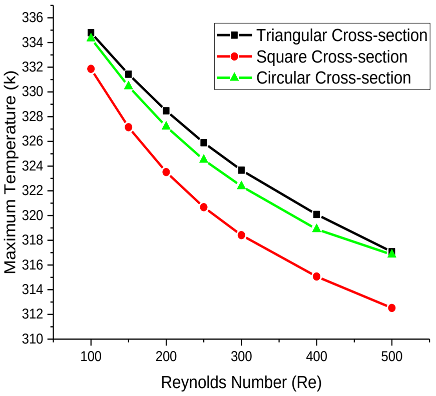


Figure 4: Maximum temperature of the cooling plate as a function of temperature

The graphs in figures 5 and 8 showed the pressure drop as a function of Reynolds number, it revealed that at higher reynolds number the diffrence between the pressure drop for respective cross-sections increases. Hence, higher pressure drop would deduce an increase of the pumping cost of the cooling flud.

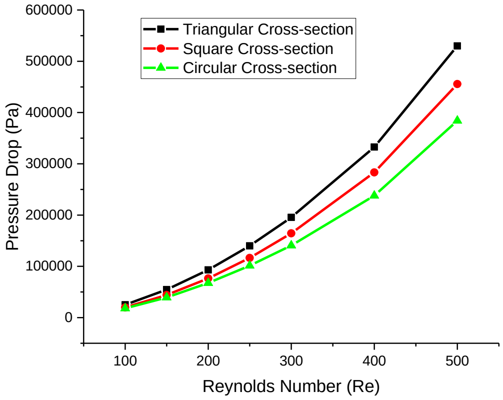
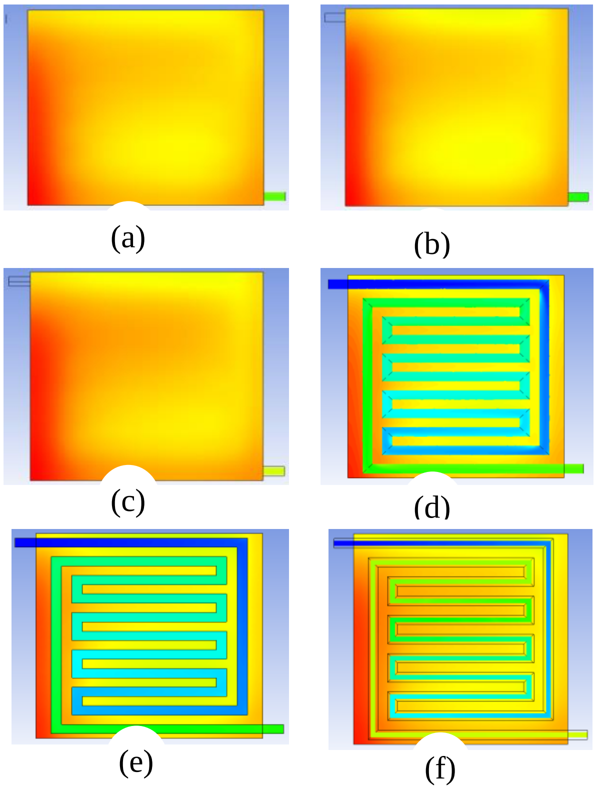


Figure 5: Pressure drop in cooling channel as a function of Reynolds number

The variation of Reynolds number against pressure drop in Fig 8 showed that the circular cross section for the new serpentine channel had the least pressure drop similar to what was observed for the conventional channel in Figure 5.

To reduce pressure drop it would be important to control the length and curvature of the fluid flow passage while ensuring uniform distribution of temperature across the plate. The figures 9 and 10 respectively described the contrast between the Maximum temperature and pressure drop for the new and the conventional channels of the cooling plates.

It was observed that the new serpentine channel had better cooling efficiency as in Figure 9 and uniform temperature distribution as in Figure 10. Also the new serpentine channel had higher pressure drop as seen in Figure 10 compared to the conventional serpentine channel.


(a)

(b)

(c)

(d)

(e)

(f)

Figure 6: Temperature profile along the top surface and temperature contour at the mid plane.

Figure (d), (e) & (f) are the temperature profile of the circular, square and triangular cross section respectively; (d), (e) & (f) are the temperature contour at the mid plane of the circular, square and triangular cross section respectively at Reynolds number of 500 for new serpentine channel.

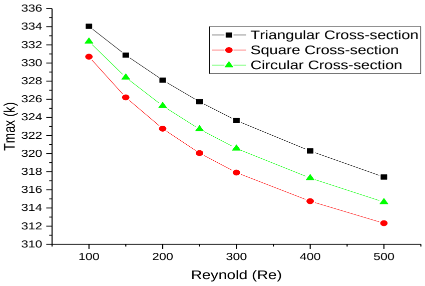


Figure 7: Maximum temperature of the cooling plate new serpentine design as a function of temperature

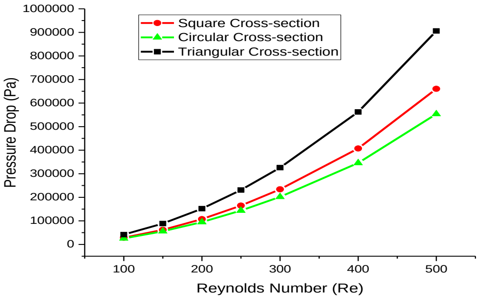


Figure 8*:* Pressure drop in cooling channel of new serpentine design as a function of Reynolds number

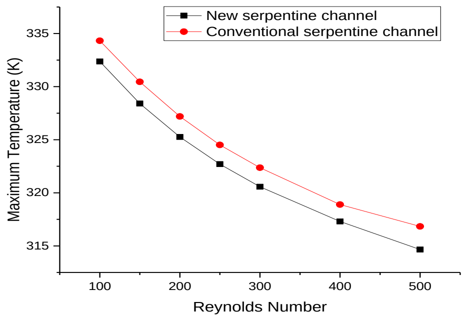


Figure 9: Maximum temperature of the circular cross sectional cooling plate for the new serpentine channel and the Conventional serpentine channel

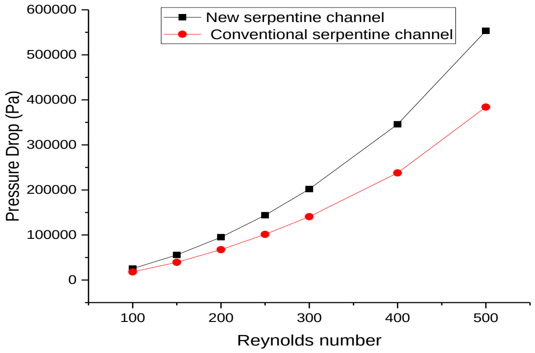


Figure 10: Pressure drop of the circular cross sectional cooling plate for the new serpentine channel and the Conventional serpentine channel

5.0 CONCLUSION

A 3D analysis of the cooling plate of a PEM fuel cell was performed using computational fluid dynamics simulation. The conventional serpentine design and a modified serpentine design was analysed by varying their cross sectional geometry at different Reynolds number. Low and even temperature/heat distribution was observed in the modified design as against the conventional serpentine design. Also, the square cross section on both the conventional design and the modified serpentine design was observed to perform better than both circular and triangular cross section. The pressure drop of the fluid channel was also calculated and the conventional serpentine design was observed to have a lower pressure drop with the circular cross section in particular having the least pressure drop in compared to the square and triangular cross section.

REFERENCES

[1] Cleghorn, S. G. S. J. C., Ren X., Springer, T. E., Wilson, M. S., Zawodzinski, C., Zawodzinski, T. A., (1977). Pem Fuel Cells for Transportation and Stationary Power Generation Applications. *Int. J. Hydrogen Energy*, 7(2), pp 23-31.

[2] Wang, Y., Chen, K. S., Mishler, J. , Cho, S. C. and Adroher, X. C. (2011). A Review of Polymer Electrolyte Membrane Fuel Cells: Technology, Applications, and Needs on Fundamental Research," *Appl. Energy*, 3(3), pp 25-35.

[3] Afshari, E., Ziaei-Rad, M., and Dehkordi, M. M. (2017). Numerical Investigation on a Novel zigzag-Shaped Flow Channel Design for Cooling Plates of PEM Fuel Cells," *J. Energy Inst.*, vol. 90 (5), pp. 752–763.

[4] Han, C. and Chen, Z. (2018). Numerical Simulation for the Effect of Vaporization Intensity in Membrane on the Performance of PEM Fuel Cell," *Numer. Heat Transf. Part A Appl.*, vol. 73(3), pp. 177–194.

[5] Li, S., Yuan, J., Xie, G, and Sundén, B. (2018). Effects of Agglomerate Model Parameters on Transport Characterization and Performance of PEM Fuel Cells," *Int. J. Hydrogen Energy*, vol. 43(17), pp. 8451–8463.

[6] Kandlikar, S. G. and Lu, Z. (2009), "Thermal Management Issues in a PEMFC Stack - A brief Review of Current Status," *Appl. Therm. Eng.*vol. 29(7), pp. 1276–1280.

[7] Nandjou F., Poirot-Crouvezier, J. P., MChandesris, M., Blachot J. F., Bonnaud, C., and Bultel, Y. (2016). Impact of Heat and Water Management on Proton Exchange Membrane Fuel Cells Degradation in Automotive Application," *J. Power Sources*, vol. 326, pp. 182–192.

[8] Zhang, G. and Kandlikar, S. G. (2012) A Critical Review of Cooling Techniques in Proton Exchange Membrane Fuel Cell Stacks," *Int. J. Hydrogen Energy*. vol. 37(3), pp. 2412–2429, 2012.

[9] Matian, M., Marquis, A., and Brandon, N. (2011). Model Based Design and Test of Cooling Plates for an Air-cooled Polymer Electrolyte Fuel Cell Stack," *Int. J. Hydrogen Energy*, vol. 36(10), pp. 6051–6066.

[10] Shahsavari, S., Desouza, A., Bahrami, M., and Kjeang, E. (2012). Thermal Analysis of Air-Cooled PEM Fuel Cells," *Int. J. Hydrogen Energy*, vol. 37(23), pp. 18261–18271.

[11] Li, S. and Sundén, B. (2018). "Numerical Study on Thermal Performance of Non-uniform Flow Channel Designs for Cooling Plates of PEM Fuel Cells," *Numer. Heat Transf. Part A Appl.*, vol. 74(1) pp. 917–930.

[12] Obayopo, S. O. (2018). CFD Modelling of Cooling Channel Geometry of PEM Fuel Cell for Enhanced Thermal Performance. Nigerian Journal of Technology, vol. 37(1), pp. 147–157.

[13] Chen, F. C., Gao, Z., Loutfy, R. O., and Hecht, M. (2003). Analysis of Optimal Heat Transfer in a PEM Fuel Cell Cooling Plate. *Journal of Power Sources*, vol. 5(4), pp. 181–188.

[14] Yu, S. H., Sohn, S., Nam, J. H., and Kim, C. J. (2009). Numerical Study to Examine the Performance of Multi-pass Serpentine Flow-Fields for Cooling Plates in Polymer Electrolyte Membrane Fuel Cells," *J. Power Sources*, vol. 194(2), pp. 697–703.

[15] Ravishankar, S. and Arul Prakash, K. (2014), "Numerical Studies on Thermal Performance of Novel Cooling Plate Designs in Polymer Electrolyte Membrane Fuel Cell Stacks," *Appl. Therm. Eng.*, vol. 66(1–2), pp. 239–251.

[16] Baek, S. M., Yu, S. H., Nam, J. H., and Kim, C. J. (2011). A Numerical Study on Uniform Cooling of Large-scale PEMFCs with Different Coolant Flow Field Designs," *Appl. Therm. Eng.*, vol. 31(8–9), pp. 1427–1434.

[17] Afshari, E. , Ziaei-Rad, M., and Shariati, Z. (2016). A Study on using Metal Foam as Coolant Fluid Distributor in the Polymer Electrolyte Membrane Fuel Cell," *Int. J. Hydrogen Energy*, vol. 41(3), pp. 1902–1912.

[18] Castelain, C., Lasbet. Y., Auvity, B., and Peerhossaini, H. (2016). Experimental Study of the Thermal Performance of Chaotic Geometries for their use in PEM Fuel Cells," *Int. J. Therm. Sci.*, vol. 101, pp. 181–192.

[19] Lasbet, Y., Auvity, B., Castelain, C., and H. Peerhossaini, H. (2007). Thermal and Hydrodynamic Performances of Chaotic Mini-channel: Application to the Fuel Cell Cooling," *Heat Transf. Eng.*, vol. 28(8–9), pp. 795–803.

[20] Kurnia, J. C., Sasmito, A. P., and A. S. Mujumdar, A. S. (2011). Numerical Investigation of Laminar Heat Transfer Performance of various Cooling Channel Designs," *Appl. Therm. Eng.*, vol. 31(6–7), pp. 1293–1304, 2011.

[21] Zakaria, I. A., Ahmad, W., and Wan, N. (2018). Numerical Analysis of Al_2O_3 Nanofluids in Serpentine Cooling Plate of PEM Fuel Cell," vol. 5(1), pp. 1–13.

FOR LEARNING AND CULTURE

ISSN 2141 - 0240